New Bid-Ask Spread Estimators from Daily High and Low Prices

Zhiyong Li and Brendan Lambe and Emmanuel Adegbite

De Montfort University

May 2017

New Bid-Ask Spread Estimators from Daily High and Low Prices

Zhiyong Li*, Brendan Lambe, Ėmmanuel Adegbite‡

forthcoming *International Review of Financial Analysis*

Abstract

Estimating trading costs in the absence of recorded data is a problem that continues to puzzle financial market researchers. We address this challenge by introducing two low frequency bid-ask spread estimators using daily high and low transaction prices.[a] The range of mid-prices is an increasing function of the sampling interval, while the bid-ask spread and the relationship between trading direction and the mid-price are not constrained by it and are therefore independent. Monte Carlo simulations and data analysis from the equity and foreign exchange markets demonstrate that these models (especially SHL2) significantly out-perform the most widely used low-frequency estimators, such as those proposed in Corwin and Schultz (2012) and most recently in Abdi and Ranaldo (2017). Using real world data we show that one of our estimators (SHL2)'s root mean square error (RMSE) is almost less than a half (even 20%) of the competitors. We illustrate how our models can be applied to deduce historical market liquidity in US, UK, Hong Kong and the Thai stock markets. Our estimator can also effectively act as a gauge for market volatility and as a measure of liquidity risk in asset pricing.

Keywords: High-low spread estimator; effective spread; transaction cost; market liquidity

JEL Classification: C02, C13, C15

*Corresponding author. Division of Finance, School of Business, University of Leicester. Email: zhiyong.li@le.ac.uk. Address: The School of Business, The University of Leicester, University Road, Leicester, LE1 7RH.

†Department of Accounting and Finance, De Montfort University, UK

‡Business School, The University of Nottingham, UK

[a] R code used in the paper is available through the link:
https://zhiyongli.weebly.com/uploads/2/4/0/1/24016076/highlowestimatorcode.zip

1 Introduction

Estimating trading costs in the absence of recorded data is a problem that continues to puzzle financial market researchers. A bid-ask spread estimator that can be used on a range of market instruments with minimum input data and which is both accurate and efficient in terms of having a low standard deviation of estimates (efficient thereafter) has become the *sine qua non* in scholarly research when true trading cost data is unavailable. Much effort is spent on arriving at an estimator that resolves this issue of opacity in historical trading cost data. New models are being introduced and ideas on how to estimate spreads have evolved considerably in the years since Roll (1984) originally formulated the concept. In this paper, we introduce two low frequency bid-ask spread estimators, these can create estimates of costs using the range between daily and two day high and low prices.

We also demonstrate how the models we propose (especially SHL2) significantly outperform existing versions, namely those introduced by Corwin and Schultz (2012) (the CS estimator hereafter), Abdi and Ranaldo (2017) (the AR estimator hereafter) and the benchmark estimator introduced in Roll (1984). We perform tests using Monte Carlo simulations, and real foreign exchange and U.S. equity market data.

Our models are designed along similar principles to the CS estimator in that it assumes that high and low prices are based on buy and sell transactions respectively. We present two models, the first, which we call our basic version uses a transaction range which is determined in part by the mid-price range and by the bid-ask spread. We posit that the mid-price range is a function of the time interval from which it is calculated. Therefore, by comparing the ranges of transaction prices from two different sampling frequencies, we can isolate the impact of the bid-ask spread. Our second model, which we refer to as our sophisticated version, builds on ideas proposed in Bleaney and Li (2016) (the BL estimator hereafter). In the model that we present, the bias that occurs as a result of feedback trading which is evident in the BL model is used to link the one and two day ranges in order to arrive at an estimation of the spread. We note that the bias that results through feedback trading is a function of the time interval. By comparing both the one and two day BL spreads we obtain our estimates of the bid-ask spread. The SHL estimator is unbiased and uses time series data for prices in addition

to using price-range. In comparison, the CS model uses only range information. The use of more relevant information will improve the estimate.

In order to analyse estimator performance, we examine the mean and standard deviation of estimated errors alongside the correlation between those and the true spreads. In addition we move beyond simply relying on one single criterion such as correlation to indicate performance. We instead show that to gauge performance on a range of indicators, such as mean, standard deviation and root mean square error (RMSE), is the optimum path to choose for researchers who are keen to attain a more accurate measure of trading costs.

The estimation of accurate bid-ask spreads has for a long time been considered a significantly important part of market microstructure theory. Bid-ask spread estimators allow researchers and practitioners to develop trading strategies that incorporate an idea of the costs attached to each transaction. In turn, this allows for a more accurate determination, of the profitability that follows on from applying . Understanding market liquidity is also important to researchers, so a precise estimation of the bid-ask spread offers a clearer picture of this market characteristic (Mancini et al. 2013; Banti et al. 2012). Another possible use for these models arises from the fact that bid-ask spreads can influence measures associated with price volatility, so scholars analysing this metric can use estimator models to arrive at an accurate measure of this (e.g. Bandi and Russell 2006).

As excessive costs are attached to accessing bid-ask spread data this has meant that researchers increasingly rely on such estimators to aid their analysis of market activity. Much research supports this approach indicating that the cheaper daily closing quoted bid-ask spread can be good proxy for the intraday spread (Holden and Jacobsen 2014, Chung and Zhang 2014 and Fong et al. 2017). Inavailability of spread data is not simply a consequence of poor research budgets, historical information on both quoted and true spreads is not always available, a strong performing estimator model is useful even to well-resourced researchers.

Bid-ask spread estimation models need to satisfy certain requirements before they become useful to researchers. Models must be accurate, efficient and it is preferable that they have low requirements on the type of data needed for computation. In order to improve on accuracy and efficiency, the signal to noise ratio becomes an important

consideration; this is because the spread (signal) is more difficult to estimate when it is considerably smaller than the mid-price volatility levels (noise). Assets with higher levels of liquidity demonstrate typically smaller spreads; however with more infrequently traded instruments, the bid-ask spreads can be quite large. Longer sampling intervals have a tendency to display higher mid-price volatility levels, therefore, the signal to noise ratio is smaller in longer sampling intervals. This leads to poorer performance in the accuracy and efficiency of estimators that rely on low sampling frequencies, this is pointed out by Bleaney and Li (2015). In addition to the need for models to be both accurate and efficient, other barriers to inquiry may inhibit a model's usefulness. Constraints on accessing data imposed by availability or cost mean that models with more modest data requirements are of greater use to researchers. For instance, Roll (1984) requires just the transaction price of assets in order to apply the estimator, whereas Huang and Stoll (1997) require both the transaction price and the order direction. Corwin and Schultz (2012) require the high low price range. Abdi and Ranaldo (2017) require the closing price and the high-low price range.

In this paper, we analyse the performance of the estimators through conducting a series of tests using both randomly generated and real data, the latter is taken from both the foreign exchange and equity markets. In most cases, both of our estimators outperform all others tested. In comparison, the CS estimator exhibits instability as it only works well for equities. The AR estimator produces estimates which are highly correlated with the true spread, while displaying a tendency to remain lower than those we estimate, it also performs poorly in terms of the root mean square error (RMSE). Simulation experiments produce a signal to noise ratio over 125000 months of generated data ranging from 0.005 to 0.387. This covers most cases which have occurred in actuality. Both of our proposed models outperform the others we test in both efficiency and accuracy. We also move beyond time series testing to investigate the cross sectional performance on a generated sample of 75000 data months, again at this level we find that our models outperform the others tested.

In the existing literature, bid-ask spread estimators are tested using the price data taken from the equities markets. An additional benefit offered by our models is that these are suitable for use both in the equities and foreign exchange markets because they are independent of the market structure. An additional contribution made through

this paper is that we verify our model's applicability by testing it directly using both FX and equity market data, other estimators also can be engaged to estimate spreads in both markets. However in their original papers which propose the approaches these are only applied to the equity markets. Subsequent papers employing these models extend the testing to the foreign exchange markets (e.g. Karnaukh et al. 2015). In this paper we run tests using data taken from both types of financial markets. In both markets the empirical tests are conducted using spreads calculated from tick by tick data as a benchmark, we find that our estimators again perform better than the other models currently available to researchers.

In presenting the significant contributions of our work, the rest of our paper is organised as follows. Section 2 discusses existing bid-ask spread estimators, while section 3 introduces our new models. In section 4 the performances of these are reported against that of the Roll, CS and AR estimators. Section 5 provides an illustrate of some applications of our estimator using equity markets while section 6 concludes.

2 Relevant bid-ask spread estimators

Spread estimator models are generally classified into one of four categories, the Roll, the LOT, the Effective Tick and the more recent High-low estimator. Each approach provides alternative methods which are based on the return autocovariance, the interval fractions in trade prices, the frequency of zero returns and the specific interval determined price range.

Roll (1984) was the first to propose a bid-ask spread estimator. This model was popularly received generating considerable interest at the time, giving rise to attempts to refine it further later by other scholars. Roll's premise was to use return autocovariance to estimate the spread. Underpinning this approach was the assumption that prices followed a random walk. It was also assumed that the closing stock price equalled its true value plus or minus half of the effective spread. The estimated spread could then be calculated as twice the square root of minus one multiplied by the autocovariance of the sample of daily returns. Some problems with this approach have been noted, for instance, the estimator produces results which can often underestimate the spread (Har-

ris 1990). To deal with this autocorrelated mid-price return bias[b], George et al. (1991) suggest modifying the original Roll estimator. Similarly, Choi et al. (1988) introduce adjustments to the model in an attempt to deal with the problem of auto correlated order directions. Stoll (1989) tackles the problem by taking the impact of inventory control and asymmetric information costs into account. To reach a general solution to the problem, Huang and Stoll (1997) incorporate each of the estimators above in one general model. However, gathering the data required to run this is a difficult process as order direction data is also required. Hasbrouck (2004, 2009) suggests that a more accurate spread can be achieved through employing Gibbs estimation. Unlike Roll's approach the computational requirements to employ Hasbrouck's estimators are considerably more intensive. The problem of normality in Hasbrouck's is addressed in Chen et al. (2016) who propose a non-parametric method to estimate the spread based on the Roll model. A further development is found in Abdi and Ranaldo (2017). which performs slightly better as the correlation between its estimates and the true spread is higher than the CS estimator, but the RMSE is not significantly better than the CS.

Another estimator used to deal with the problem of the spread opacity due to the inavailability of data is proposed by Lesmond et al. (1999). Otherwise known as the LOT model, an effective spread is calculated by considering the fraction of returns which are different from zero. This model has not quite reached the popularity levels of the Roll model as comparatively it tends not to perform as well in empirical testing (Corwin and Schultz 2012). Holden (2009) and Goyenko et al. (2009) put forward a more sophisticated approach which estimates spreads using effective tick measures based on the phenomenon of price clustering, a term describing the tendency for trade prices to occur most frequently on rounder price increments. However, results produced following testing on an extensive FX market data sample by Karnaukh et al. (2015) show that the LOT and effective tick estimators display only a weak relationship with true spreads.

The high-low spread estimator introduced by Corwin and Schultz (2012) adds new power to the toolkit of estimators. Despite being relatively new, it is used extensively in recent literature as testing shows that it satisfies the estimator requirements to a greater extent than previous innovations (Corwin and Schultz 2012, Holden and Jacobsen 2014

[b]The bias arises as the assumption that returns are random is not satisfied.

and Karnaukh et al. 2015).

3 The High-low estimators

The general structure of both of our estimators can be expressed in the following equation:

$$\frac{E\left(\sqrt{2}\cdot X_{daily} - X_{twoday}\right)}{\sqrt{2}-1} \tag{1}$$

Where X is the price range or the estimated spread by the Bleaney and Li (2016) estimator. The innovations we introduce for both the basic and sophisticated models are discussed in the following subsections. The combination of basic and the sophisticated high-low estimator introduced in the next section can in some circumstances outperform a single estimator; this is because the difference in the estimated errors from each can offset each other to some extent.

3.1 The basic high-low estimator (the BHL model)

Up to the introduction of our model, the CS estimator had been the best performer out of the array of models available to researchers, however it suffers from some bias owing to its non-linear structure[c]. Our basic high-low estimator is similar to the CS model, but rather than having a quadratic structure the BHL is linear, ensuring the unbiasedness of its estimates.

When estimating a model using the high and low transaction prices, the first characteristic that we can note is that the range increases as the time interval widens and as the mid-price volatility grows in proportion. The other factor contributing to the range is the bid-ask spread, however this is independent of the time interval. Therefore, it is possible to extract the bid-ask spread by calculating the difference between the high and low transaction prices, whilst considering inconsistencies that may arise as a result of volatility.

In order to accomplish this, we assume that the mid-price, denoted as M_t, follows a one-dimensional Wiener process. The link between the unobserved mid-price and the

[c]This is detailed in the error analysis available in Corwin and Schultz (2012) and Bleaney and Li (2015).

observed transaction price (s_t) is given through the following equation.

$$s_t = M_t + \frac{SP}{2} \cdot BS_t \tag{2}$$

Where BS_t is the trade indicator showing 1 (-1) for a buyer (seller) initiated trade. The relationship between the daily high mid-price (H_t^M) and the daily high transaction price (H_t^T) as well as the link between the daily low mid-price (L_t^M) and the daily low transaction price (L_t^T) are demonstrated in the following set of equations:

$$\begin{aligned} H_t^T &= H_t^M + \tfrac{SP}{2} \cdot BS_t \quad L_t^T = L_t^M + \tfrac{SP}{2} \cdot BS_t \\ TH_t^T &= TH_t^M + \tfrac{SP}{2} \cdot BS_t \quad TL_t^T = TL_t^M + \tfrac{SP}{2} \cdot BS_t \end{aligned} \tag{3}$$

Where T and M represent the transaction and mid-price respectively. TH and TL denote the high and low prices over a two day window.

We can eliminate the need to establish order direction by assuming that the highest (lowest) prices are buy (sell) orders. Formally, it can be represented as:

$$\mathrm{BS_t} = \begin{cases} 1 \quad if \ \ \mathrm{s_t} = H_t^T \\ -1 \ \ if \ \ \mathrm{s_t} = L_t^T \end{cases} \tag{4}$$

The daily and two-day ranges of transaction prices represent the difference between the highest and lowest prices. Formally, taking equations (3) and (4) into account, these ranges are given as:

$$\begin{aligned} Range_{t,daily}^T &= H_t^T - L_t^T \\ &= \left(H_t^M + \tfrac{SP}{2} \cdot BS_t\right) - \left(L_t^M + \tfrac{SP}{2} \cdot BS_t\right) \\ &= \left(H_t^M + \tfrac{SP}{2}\right) - \left(L_t^M - \tfrac{SP}{2}\right) \\ &= \left(H_t^M - L_t^M\right) + SP \\ &= Range_{t,daily}^M + SP \end{aligned} \tag{5}$$

$$\begin{aligned} Range_{t,twoday}^T &= TH_t^T - TL_t^T \\ &= \left(TH_t^M + \tfrac{SP}{2} \cdot BS_t\right) - \left(TL_t^M + \tfrac{SP}{2} \cdot BS_t\right) \\ &= \left(TH_t^M + \tfrac{SP}{2}\right) - \left(TL_t^M - \tfrac{SP}{2}\right) \\ &= \left(TH_t^M - TL_t^M\right) + SP \\ &= Range_{t,twoday}^M + SP \end{aligned} \tag{6}$$

Where $Range_{t,daily}^T$ and $Range_{t,twoday}^T$ are daily and two-day ranges respectively. The equations above demonstrate our earlier suggestion that the range of transaction prices

is influenced by volatility in both the mid-price and the bid-ask spread. Taking expectations of both sides, the equations become[d]:

$$E\left(Range^{T}_{daily}\right) = E\left(Range^{M}_{daily}\right) + SP \tag{7}$$

$$E\left(Range^{T}_{twoday}\right) = E\left(Range^{M}_{twoday}\right) + SP \tag{8}$$

The left hand sides of Equations (7) and (8) can be calculated from observed transaction prices. With the unobserved terms, the expected ranges of daily and two-day mid-prices can be eliminated, allowing us to extract the bid-ask spread. Parkinson (1980) shows that if the mid-price follows a one-dimensional Wiener process, its expected range is an increasing function of the sampling time interval and its diffusion. A long sampling time interval or large diffusion will lead to a wider range. Formally, the expectation of the range of mid-prices can be calculated through the following equation:

$$E\left(Range^{M}\right) = \sqrt{\frac{8D \cdot ti}{\pi}} \tag{9}$$

Where D is the diffusion of mid-prices in a unit time interval (ti). If this period is one day, the expectations for daily and two-day ranges are given through the following equations:

$$E\left(Range^{M}_{daily}\right) = \sqrt{\frac{8D}{\pi}} \tag{10}$$

$$E\left(Range^{M}_{twoday}\right) = \sqrt{\frac{8D}{\pi}} \cdot \sqrt{2} \tag{11}$$

Therefore, the expectation is that the two-day range is $\sqrt{2}$ times that of the daily range. Formally, the relationship is expressed through the following equation:

$$E\left(Range^{M}_{twoday}\right) = \sqrt{2} \cdot E\left(Range^{M}_{daily}\right) \tag{12}$$

[d] Equations (7) and (8) demonstrate the key difference between the BHL and CS estimator models. Unlike BHL which uses the first moment, the CS estimator uses the second moment in both equations as follows.

$$\left[E\left(Range^{T}_{daily}\right)\right]^2 = \left[E\left(Range^{M}_{daily}\right) + SP\right]^2$$

$$\left[E\left(Range^{T}_{two}\right)\right]^2 = \left[E\left(Range^{M}_{two}\right) + SP\right]^2$$

From Equations (7), (8) and (12), we can solve for the bid-ask spread (SP), because we have three equations and three unknown variables. We solve Equation (8) through deducting $\sqrt{2}$ times each side of Equation (7):

$$\begin{aligned} & E\left(Range^{T}_{twoday}\right) - \sqrt{2} \cdot E\left(Range^{T}_{daily}\right) \\ & = E\left(Range^{M}_{twoday}\right) + SP - \sqrt{2} \cdot \left[E\left(Range^{M}_{daily}\right) + SP\right] \end{aligned} \tag{13}$$

When we substitute Equation (12) into (13), and rearrange the yields, the estimate of the bid-ask spread becomes:

$$SP = \frac{E\left[\sqrt{2} \cdot \left(Range^{T}_{daily}\right) - \left(Range^{T}_{twoday}\right)\right]}{\left(\sqrt{2}-1\right)} \tag{14}$$

Equation (14) represents the basic estimator which we propose in this paper (BHL hereafter); this is an expectation of the linear function of the daily and two-day high and low transaction prices[e]. One of its key features is that it is unbiased and easy to compute. It outperforms the CS estimator because it produces an unbiased result while remaining linear. Using BHL, it is possible to increase the number of observations in order to obtain a better estimate of the spread. The reason is that statistical errors and noise can be eliminated from large sample sizes; this is not the case for non-linear estimators (Bleaney and Li 2015). Furthermore, the estimates remain stable across a variety of sampling periods. This suggests that when higher sampling frequency data becomes available we can use it to obtain a better estimate because, as Bleaney and Li (2015) suggest, the noise (the price volatility) is relatively low in comparison with the bid-ask spread.

Similar to the CS estimator, we can also estimate the daily diffusion, which is expressed as D, using the same process. This is represented in the following equation:

$$\begin{aligned} E\left(Range^{M}_{daily}\right) &= \frac{E\left(Range^{T}_{twoday}\right) - E\left(Range^{T}_{daily}\right)}{\left(\sqrt{2}-1\right)} = \sqrt{\frac{8D}{\pi}} \\ D &= \frac{\pi}{8}\left[\frac{E\left(Range^{T}_{twoday}\right) - E\left(Range^{T}_{daily}\right)}{\left(\sqrt{2}-1\right)}\right]^{2} \end{aligned} \tag{15}$$

[e]When the when the sample size is small, or the mid-price volatility is big, the BHL could underestimate the spread or even have negative spread. Formal prove can be provided upon request.

3.2 The sophisticated high-low estimator

Our sophisticated estimator (the SHL model hereafter) introduces innovations to the design proposed by Bleaney and Li (2016) whilst sharing the same structure and settings with BHL. The differentiators are that the SHL estimator is unbiased and incorporates more features of high and low data than BHL, this then provides more accurate estimates than those offered through BHL. The BL estimator is distinctive in that it outperforms Roll (1984), Huang and Stoll (1997), Corwin and Schultz (2012) and Hasbrouck (2009) estimators, following extensive testing. The BL model requires trade direction and transaction price, due to the fact that trade direction data is unavailable to most researchers, we innovate through the SHL in order to remove this constraining requirement. The data requirements are therefore less demanding than the BL model using only high and low prices.

SHL introduces the assumption that the highest prices recorded daily are ask-prices and the lowest are bid-prices. Through this assumption we can lower the data requirements for the model and allow the estimator to operate using only the highest and lowest transaction prices in the estimation window.

In a similar manner to Bleaney and Li (2016), we assume that we have random conjectures of the true bid-ask spread. We let set A be a set of all conjectures where the symbol $\sim$ represents conjectural values.

$$A = \left\{ \widetilde{SP}_1, \widetilde{SP}_2, \cdots, \widetilde{SP}_n \right\} \tag{16}$$

At this stage, we do not know which element in set A is the true spread. Through taking the following steps, we would be able to find it. First, we would calculate a series of conjectural mid-price returns according to each element (a conjectural spread) in set A using equation (2). Formally, the conjectural mid-price return is given as follows,

$$\widetilde{M}_t = s_t - \frac{\widetilde{SP}_t}{2} \cdot BS_t \tag{17}$$

Second, we calculate the variance of conjectural mid-price returns for each conjectural series.

B denotes a set of variances of conjectural mid-price returns, the conjecture being that the true spread is taken to be:

$$B = \{Var_1, Var_2, \cdots, Var_n\} \tag{18}$$

Where

$$Var_i = Var\left[\Delta\widetilde{M}(\widetilde{SP}_i)_t\right] \tag{19}$$

Third, based on these settings we can find the true spread among the conjectures by find the biggest relevant variance.

Figure 2 outlines the reasoning underpinning this process where for the purposes of economy we hold that the mid-price is fixed, , while mid-prices following a random walk will not affect the derivation of the model. The conjectural spread ($\widetilde{SP}_i$) is less than the true spread. This allows us to estimate the conjectural mid-price $\widetilde{M}$; this is represented by the dotted line in Figure 2, and the true mid and transaction prices are both represented by unbroken lines. Also in Figure 2, A and B denote observed ask and bid prices, whereas M is the unobserved true mid-price.

At any one point we can only observe one price, which is either the bid or ask. In Figure 2, three periods are displayed. In the period labelled $t-2$, the bid price is recorded and in period labelled $t-1$, the ask price is observed. In period $t-2$, the conjectural spread is lower than the true spread and the conjectural mid-price error is -0.5Ω , which is less than the true value. In period $t-1$, the conjectural mid-price error is 0.5Ω, therefore this is greater than the true one. In the intervening period between $t-2$ and $t-1$, the direction of the trade shifts from sell to buy, and because of the conjectural error, we overestimate the mid-price return, formally we express this as:

$$\Delta\widetilde{M}_{t-1} = \Delta M_{t-1} + \Omega = \Omega \tag{20}$$

In Figure 2, the hypothetical example shows that the variance of mid-price returns equates to zero because returns remain fixed. However the variance of conjectured mid-price returns is greater than zero. The reason for this is that in the case where the spread is underestimated, the conjectured mid-price fluctuates more than its true counterparts.

[Insert Figure 2 here]

Formally, we propose the following:

Proposition 3.1 *When the components of the spread do not include feedback trading, inventory control or asymmetric information, we can consider that the spread and its estimates, and thus the estimated errors, are either serially independent or fixed. If an estimate of the spread*

$\widetilde{SP}_i \in A$ corresponds to $Var_i = max(B)$, it equals the true spread which is then denoted as: $\widetilde{SP}_i = SP$.

Proof The full proof is given in the appendix. The variance of the conjectures of mid-price returns is:

$$\begin{aligned} Var_i &= Var\left[\Delta\widetilde{M}_t\right] \\ &= E\left\{\left[\Delta\widetilde{M}_t - E\left(\Delta\widetilde{M}_t\right)\right]^2\right\} \end{aligned} \tag{21}$$

We will assume that the expectation of the value of the conjectural mid-price is zero. Thus, the equation above can be rewritten as:

$$\begin{aligned} Var_i &= Var\left(\Delta\widetilde{M}_t\right) \\ &= E\left(\Delta\widetilde{M}_t{}^2\right) \\ &= E\left[\left(\Delta M_t + \tfrac{1}{2}\Omega\, BS_t - \tfrac{1}{2}\Omega\, BS_{t-1}\right)^2\right] \end{aligned} \tag{22}$$

Where Ω denotes the conjectural error which represents the difference between the conjectural mid-price and the true mid-price, alternately expressed as the difference between the conjectural spread and its true value. Formally, Ω is given as:

$$\Omega = \Delta\widetilde{M}_t - \Delta M_t = \widetilde{SP}_i - SP \tag{23}$$

The assumptions of this proposition imply that BS is independent of ΔM at all observation points, therefore many of the terms in (22) such as $E(\Delta M_t\, BS_t)$ and $E(\Delta M_{t-1}\, BS_t)$ equate to zero. The variable BS is a binary variable (1 or -1), thus $E\left(BS_{t-1}{}^2\right) = 1$. Finally we obtain:

$$\begin{aligned} Var_i &= Var\left(\Delta\widetilde{M}_t\right) \\ &= E\left(\Delta\widetilde{M}_t{}^2\right) \\ &= E\left[\left(\Delta M_t + \tfrac{1}{2}\Omega\, BS_t - \tfrac{1}{2}\Omega\, BS_{t-1}\right)\left(\Delta M_t + \tfrac{1}{2}\Omega\, BS_t - \tfrac{1}{2}\Omega\, BS_{t-1}\right)\right] \\ &= E\left(\Delta M_t{}^2 + \tfrac{1}{2}\,\Omega^2\right) \end{aligned} \tag{24}$$

The final step of Equation (24) given above is the quadratic polynomial of the expectation of the error of the conjecture. For a given series, the first term $E(\Delta M_t^2)$ is a constant. We can surmise directly from this that when the error is zero (i.e. $\Omega = 0$), the second term $\frac{1}{2}\Omega^2$ is zero. Furthermore, when $\Omega = 0$, there is a global extreme for the right hand side polynomial in the final step, and symmetrically, the left hand side

of the equation $Var_i = Var\left(\Delta\widetilde{M}_t\right)$ is also at the extreme value. Formally this can be expressed as:

$$\underset{\Omega}{arg\ max}\, Var\left(\Delta\widetilde{M}_t\right) = 0 \tag{25}$$

When the conjectural error is zero, the conjectural spread becomes the true spread:

$$\widetilde{SP}_i = SP + \Omega = SP \tag{26}$$

Therefore the conjectural spread which maximises the covariance equals the true spread.

$$\underset{\widetilde{SP}_i \in A}{arg\ max}\, Var\left(\Delta\widetilde{M}_t\right) = SP \tag{27}$$

Q.E.D.

According to the abovementioned proposition, we find that the true spread maximises the variance of conjectural mid-price returns and can be expressed as follows:

$$\begin{aligned} &Var\left(\Delta\widetilde{M}_t\right) = E\left(\Delta\widetilde{M}_t{}^2\right) \\ &= E\left[\left(\Delta s_t - \tfrac{\widetilde{SP}}{2}\Delta BS_t\right)^2\right] \\ &= E\left(\Delta s_t{}^2\right) - \widetilde{SP}\cdot E\left(\Delta s_t\, \Delta \mathrm{BS}_t\right) + \tfrac{\widetilde{SP}^2}{4}\cdot E\left(\Delta BS_t{}^2\right) \end{aligned} \tag{28}$$

Using first order conditioning, we find that the estimated spread satisfies the following equation:

$$-E\left(\Delta s_t\, \Delta BS_t\right) + \frac{1}{2}\widetilde{SP}\cdot E\left(\Delta BS_t{}^2\right) = 0 \tag{29}$$

$$SP = \widetilde{SP} = \frac{2E\left(\Delta s_t \Delta BS_t\right)}{E\left(\Delta BS_t{}^2\right)} \tag{30}$$

Equation (30) is now the variance version of the BL estimator, thereby reflecting one of the suggested innovations that we propose in this paper.

Equation (30) requires order directions to estimate the spread. In order to allow it to become operational without order directions, we must introduce the following processes. On each day, we pick either the high or low price randomly to create a trail series of prices (s_t) and use Equation (4) to determine the order direction: a buy order when s_t is the high price and a sell order when s_t is the low price. We can then calculate the estimated spread using Equation (30). In the same manner as Corwin and Schultz (2012), we calculate an estimate of spread using the two-day high and low prices SP_{twoday}.

However, Equation (4) creates the link between order flow and price when only high and low prices are used. When the covariance between order directions and mid-price returns is non-zero, the BL estimator is biased and the error is expressed as $E\left(BS_t \cdot \Delta M_t\right)$ [f]. Therefore, when high and low prices and relevant order directions are used, the BL estimator significantly overestimates the spread. Formally, the estimated spread is the true spread plus the erros:

$$\begin{aligned} SP_{daily} &= SP + \underbrace{E\left(BS_{daily} \cdot \Delta M_{daily}\right)}_{error} \\ SP_{twoday} &= SP + \underbrace{E\left(BS_{twoday} \cdot \Delta M_{twoday}\right)}_{error} \end{aligned} \tag{31}$$

Where the subscripts "daily" and "twoday" represent the sampling frequencies. SP_{daily} and SP_{twoday} are the BL estimates using daily and two-day high low data respectively.

It is invariably the case that the estimated spread will contain errors, these nevertheless can be offset if we compare the estimates using daily and two-day data ($E(\mathrm{SP}_{daily})$ and $E(\mathrm{SP}_{twoday})$). This is because it is possible to predict the relationship between errors from daily and two-day estimates.

Following the discussion in the previous section, the relationship between daily and two-day ranges can be used to eliminate the estimated error above. The errors are in fact half of expected ranges of daily and two-day ranges and are expressed as follows (proof can be found in the appendix):

$$\begin{aligned} E\left(BS_{daily} \cdot \Delta \mathrm{M}_{daily}\right) &= \tfrac{1}{2}E\left(\mathrm{H}_t^M - \mathrm{L}_t^M\right) = \tfrac{1}{2}E\left(Range_{daily}^M\right) \\ E\left(BS_{twoday} \cdot \Delta M_{twoday}\right) &= \tfrac{1}{2}E\left(TH_t^M - TL_t^M\right) = \tfrac{1}{2}E\left(Range_{twoday}^M\right) \end{aligned} \tag{32}$$

Following steps similar to the process outlined in section 3.1, we substitute Equations (32) and (12) into Equation (31), the rearrangement yields the spread for one trial series. This trial estimate is given as:

$$SP_{onetrial} = \frac{\sqrt{2}\,SP_{daily} - SP_{twoday}}{\sqrt{2}-1} \tag{33}$$

We repeat the trail series creation and estimation process a number of times, the mean of these estimates becomes our estimation of the spread. Although this process is com-

[f] The feedback trading bias is discussed in Bleaney and Li (2016). Proofs are given in the appendix of this paper.

putationally intensive, this makes little practical difference given the power of the current stock of computers available to researchers.

$$SP = \frac{1}{N}\sum_{i=1}^{N}(SP_{onetrial})_i = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sqrt{2}\,SP_{daily} - SP_{twoday}}{\sqrt{2}-1}\right)_i \tag{34}$$

Equation (34) is the sophisticated high low estimator (SHL), where N is a large number, for example 1000. Theoretically, SHL should produce more accurate results than its BL and CS counterparts. Unlike the BL model, the SHL estimator will not be influenced by feedback trading and the estimates produced will be unbiased. Furthermore, SHL uses time series data for Δs and ΔBS in addition to using price-range. In comparison, the CS model uses only range information. The use of more relevant information will improve both the accuracy and efficiency of the estimate.

When the ratio of the spread to the standard deviation of mid-price is small, some trail estimates in Equation (33) could be negative. In practice it is common for many estimators to generate negative or undefined outcomes. We see this for example in Roll's, Corwin and Schultz's estimates.Despite the frequency of occurrence it is difficult to explain the negative results. The bid-ask spread represents the trading costs. All the models introduced thus far demonstrate the capacity to produce negative spreads; it is economically meaningless to suggest that the trade costs can ever be negative. Therefore, we introduce a mechanism to circumvent the dilemma of the appearance of negative spreads (through our SHL model). When the trial estimates are negative for Equation (33) we restrict the number to zero. Formally, the second version of SHL (thereafter SHL2) is given as follows,

$$SP = \frac{1}{N}\sum_{i=1}^{N} max\left[\left(\frac{\sqrt{2}\,SP_{daily} - SP_{twoday}}{\sqrt{2}-1}\right)_i, 0\right] \tag{35}$$

We also consider the possibility that SHL2 may overestimate the spread, in the next section we conduct simulation experiments and the results of these show that it will not be an issue when the ratio becomes larger.

We can also estimate the daily diffusion, which is expressed as D, from Equations (31) and (32) and using the same process. This is represented in the following equa-

tions:

$$\begin{aligned} &SP_{twoday} - SP_{daily} = E\left(BS_{twoday} \cdot \Delta M_{twoday}\right) - E\left(BS_{daily} \cdot \Delta M_{daily}\right) \\ &= \tfrac{1}{2}E\left(Range^{M}_{twoday}\right) - \tfrac{1}{2}E\left(Range^{M}_{daily}\right) \\ &= \tfrac{(\sqrt{2}-1)}{2}\sqrt{\tfrac{8D}{\pi}} \end{aligned} \tag{36}$$

The rearrangement of the equation above yields an expression of daily diffusion as follows:

$$D = \frac{\pi}{2} \cdot \left[E\left(\frac{SP_{twoday} - SP_{daily}}{\sqrt{2}-1}\right)\right]^2 \tag{37}$$

4 Comparison of the estimators

In this section, we examine the performance of a range of estimators. Using empirical tests, we gauge how BHL, SHL, Roll, CS and AR perform in addition to a number of equally weighted combinative models. Currently, the range of estimators available to researchers is wide, but we focus on these models for several reasons. The first is that we wish to contrast the performance of our proposed models (BHL and SHL) with that of the best performing estimator available, the CS model. Corwin and Schultz (2012) demonstrate that the CS model outperforms all other low frequency estimators in terms of accuracy and efficiency. Holden and Jacobsen (2014) and Karnaukh et al. (2015) also show similar results to the model originators. We also choose the Roll model as this has traditionally been the benchmark for estimator performance. Researchers less familiar with the relatively recent CS model can understand how our models perform in comparison. Finally, we select the model proposed in Abdi and Ranaldo (2017) because it is the latest development of the spread estimator and is directly related to both the Roll and CS models. Our motivation behind including the combinatory models relates to the tendency for some estimators to over(under)estimate the spread. Combinations of estimators have been shown in (Holden 2009)[g] to perform better in terms of accuracy. The data we use to test each of the models is tick by tick equity prices and foreign exchange rates; these are sourced from TAQ and Hotspot respectively. In addition to using real world data testing, simulation experiments were also carried out. Our findings show that in general our BHL and SHL estimators outperform all other estimators included in the study.

[g]Combination models tested in Holden (2009) and used here for testing are explained in Table 1.

4.1 Comparison strategy

Following testing on each of the estimators, the results are reported using average relative estimated errors together with their root mean square and standard deviation values. Formally, the relative error is defined as follows:

$$Rel - Err = \frac{Estimates - Spread}{Spread} \tag{38}$$

The average relative error (Rel-Error-Mean) reports the mean difference between the estimated and true spread, indicating where possible bias may exist in the estimators. When Rel-Error-Mean is positive (negative) it suggests that the models over (under) estimate the spread. Good estimates are those with 'close-to-zero' relative error averages. Formally, this is presented as:

$$Rel - Err - Mean = E\left(Rel - Err\right) \tag{39}$$

The standard deviation of the relative estimated errors (Rel-Err-Std) is also reported and provides a measure for the efficiency of the estimates. Good estimates have low Rel-Err-Stds. Formally, this is expressed as:

$$Rel - Err - Std = Std.Dev\left(Rel - Err\right) \tag{40}$$

Finally, the RMSE is the most widely used criteria by which to judge the performance of the estimators. Therefore we follow this trend in analysing how our models perform. Formally, RMSE is given as:

$$RMSE = \sqrt{E\left[\left(Rel - Err\right)^2\right]} \tag{41}$$

4.2 Simulation experiments

In this section we report the results of a number of simulations designed to test the relative strength of each measure. The key settings for the simulation are almost the same as those used in Corwin and Schultz (2012). Ours offers a more fitting contextual scenario as it reflects the continuous operation of the 24 hour forex market[h]. We find

[h]The difference in our simulation is that we assume a 24 hours per day schedule rather than Corwin and Schultz (2012) testing assumption that proposed 390 minutes per day, this is based on stock market opening hours.

that the estimators proposed in this paper outperform the other models in terms of accuracy and efficiency. Simulation experiments are widely used in literature to examine and to compare various estimators (e.g. Corwin and Schultz 2012, Bleaney and Li 2015, 2016,Karnaukh et al. 2015,Abdi and Ranaldo 2017). Compared to the real data, the statistical properties of estimators can be extracted using a large number simulation experiments. One also could identify the factors that influence the performance of estimators, which help researchers to choose from various estimators according to their purpose.

4.2.1 Estimation under various 'signal to noise' ratios

It is difficult for an estimator to isolate the bid-ask spread from transaction prices when the volatility of mid-prices is relatively large. We test the performance of the estimators under various 'signal to noise' values which are the ratios of the spread to the standard deviation of the mid-price. The 'signal to noise' ratio is low for heavily traded equities and major currency pairs because the liquidity levels are consistently high, and the assumption is that mid-prices and order directions are random[i].

We allow the standard deviation of one-minute mid-price returns to be 0.005 (about 0.19 daily). We consider six bid-ask spreads ranging from 0.001 to 0.3. The 'signal to noise' ratio extends from 0.00527 to 1.58 on a daily basis. In comparison, Corwin and Schultz (2012) test their model using the ratios which begin at 0.167 and end at 3.33; therefore the performance hurdles we employ to evaluate our estimators are more difficult to overcome.

Our simulation experiments are therefore more challenging and mirror real market conditions. For example, assuming that there are 20 trading days in a month, we compare the estimates of 25000 months. Formally, the data generation system is given

[i]The estimators do not depend on the random walk assumption. In appendix 7.5, we show that the estimators' performance is not significantly influenced by the auto-correlated mid-price returns, which is the same as Bleaney and Li (2015)

as:

$$
\begin{aligned}
& s_t = M_t + \tfrac{SP}{2} \cdot BS_t \\
& BS_t \sim B(1, 0.5) \\
& \Delta M_t \sim N(0, 0.05) \ (one - minute) \\
& SP = \begin{cases} 0.001 \ report\ online\ for\ brevity \\ 0.002 \ report\ online\ for\ brevity \\ 0.006 \ report\ online\ for\ brevity \\ 0.010 \ report\ online\ for\ brevity \\ 0.030 \end{cases}
\end{aligned}
$$

Table2 reports the testing using various versions of our BHL, SHL and CS models[j]. In general, those estimates are more accurate and efficient from the top left to the bottom right, as the ratio (*True spread/Midstd*) increases (from 0.00527 to 0.387) the number of observations increases[k]; this is consistent with findings of Bleaney and Li (2015, 2016). The right panel of Table 2, reporting four-hour cases, demonstrates the fact that the estimators can also be used for different sampling frequencies, similar testing parameters were conducted by the online appendix of Corwin and Schultz (2012). By setting negative trials and results to be zero, the BH3, SHL2, CS3, AR and Roll estimators demonstrate significant bias. For example, the relative error of SHL2 is 74.97% and those of BH3, CS3 and AR are 330%, 234% and 155% respectively when the ratio is 0.158 (The left panel of Table 2). If the ratio is 0.387 (The right panel of Table 2), the relative error of SHL2 is -0.593% which is close to zero, and therefore demonstrates the power of the model. For the other estimators, the relative errors of SHL1, BHL1 and BHL2 are less than 5% when the ratio is greater than 0.0258. According to the second column of Table 4, the average ranking of all simulation experiments in section 4.2.1 suggests that the combination of SHL2 and BHL1 is the best performing estimator. In terms of the performance of single rather than combined estimators, SHL2 offers the best results and is the second best performer from the entire array of models.

[Insert Table 2 here]

[j] See the caption of Table 2 for full details of the versions of the estimators used.

[k] Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation. We also test the cases of full sample and the case where the trimming is at the 0.05% and 2% level, the results produced are similar.

We also test the asymptotic properties of the estimators in this section. The simulation data used in the right panel (four-hour case) in Table (2) is selected to construct the asymptotic distributions of the estimators. In existing tests, for the results are given in Table (2), there are 120 observations in a group per month. To demonstrate the asymptotic characteristics of the results produced through the simulation we consider five cases where there are 6, 12, 50, 100 and 200 observations for each group. We then generate estimated spreads according to each case (note there are 15000 groups in each). The distributions of the estimates in the groups in each case are shown in Figure 3. These figures suggest that as the number of observations in a group increases, the estimates converge toward the true spread gradually with the exception of estimations produced by the ROLL and AR models. We see that this holds true for our SHL2 model, where as we increase the number of observations, the estimates converge to the true value. In contrast, as observations are increased, AR's estimates move away from the true spread and converge around the mean value. [Insert Figures 3 here]

4.2.2 Cross-sectional properties of the estimators

In this section, the cross-sectional properties of the estimators are examined. In contrast to the previous section, the bid-ask spreads are assumed to vary each month and are evenly distributed from 0.002 to 0.0177. We also break the full sample into five groups according to the mean of the bid-ask spread. Thus, we can examine the cross-sectional performances of the estimators across the five ranges of spread. The other parameters in the data generation process are the same as in the previous section.

Formally, the data generation system is given as:

$$
\begin{aligned}
& s_t = M_t + \tfrac{SP}{2} \cdot BS_t \\
& BS_t \sim B(1, 0.5) \\
& \Delta M_t \sim N(0, 0.05) \ (one - minute) \\
& SP = \begin{cases} from\ 0.001\ to\ 0.00513\ report\ online\ for\ brevity \\ from\ 0.00513\ to\ 0.00829\ report\ online\ for\ brevity \\ from\ 0.00829\ to\ 0.0114\ report\ online\ for\ brevity \\ from\ 0.0114\ to\ 0.0146\ report\ online\ for\ brevity \\ from\ 0.0146\ to\ 0.0177 \end{cases}
\end{aligned}
\tag{42}
$$

The results of the simulation experiments are reported in Table 3. The correlations between the true and estimated spreads are also reported. Table 3 reports the pooled results while the other panels are represented in the equation above according to each grouping of spreads. In the case of spreads where the range is between 0.01 and 0.03, the correlations reported are quite weak.

In the pooled case, although CS3 has a slightly stronger correlation than SHL2 with values of 0.136 and 0.127 respectively, it reports a much higher value for RMSE at 12.32 that the SHL2 value which is 5.51. In terms of correlation, the best performers are CS3, SHL2 and BHL3. From the third column of Table 4, it is evident from the average ranking of all simulation experiments that the combination of SHL2 and BHL1 and that of SHL2 and CS2 are the best performing estimators. For single models, SHL2 shows the best performance and is placed third in rank overall.

Table 4 shows a summarised average ranking for all simulations. According to the first column, the average ranking of all cases of simulation experiments suggests that the combination of SHL2 and BHL1 is the best performing estimator. SHL2 is the best performing single estimator and takes second place overall. The other combinations outperform the other single estimators. The remaining alternate versions of our new models (BHL1, BHL2, SHL1) perform better than all the versions of the CS and Roll estimators.

[Insert Tables 3 to 4 here]

4.3 Comparisons in foreign exchange markets

In this section, we use our chosen estimators to gauge historical spreads for the foreign exchange markets. We test the estimators using the prices and effective spreads of 23 Currency pairs in a sample dating from December 2015 to August 2016, this data is taken from Hotspot. In testing on both currency samples our sophisticated high-low estimator outperforms all others employed in the test.

4.3.1 Hotspot 23 currency pairs Dec 2015-Aug 2016

In this section, we evaluate the performance of the estimators using daily high and low prices and the effective time-weighted bid-ask spread data of 23 currency pairs

sourced from Hotspot. Results are reported in Table 5. Hotspot is a large electronic communication network (ECN) platform for foreign exchange transactions conducted worldwide. We extract quotes and transaction data similar to that taken from the TAQ database. Trade volume weighted effective spreads are calculated for each pair over time the sample period begins in December 2015 and ends in August 2016. Spreads are arrived at through the matching of quote and transaction data. The trade volume weighted effective spread can be formally expressed as:

$$\begin{aligned} &2 \cdot (s_t - M_{t-1}) \quad for\ buyer\ initiated\ trades \\ &2 \cdot (M_{t-1} - s_t) \quad for\ seller\ initiated\ trades \end{aligned} \tag{43}$$

In order to reduce the possibility of errors in the data we eliminate outliers[1] and negative effective spreads. Table 5 displays the results of the pooled case where the 23 currencies over the entire sample period are examined. SHL2 is the best performing model in terms of RMSE, although its estimated error is high (230%) but is similar to the others models. The standard deviation of SHL2 is the lowest (3.2) among the estimators tested, where each model either significantly over or under-estimates the spread. Although CS2 has the lowest average estimated error (14%), its standard deviation of 6.123 is large relative to the others tested. The currency-month pooled correlation coefficients for the estimated and true spreads for BHL3 and CS3 are 0.95 and 0.94 respectively; these are much higher than the other models tested. However, their relative errors and standard deviations are greater in magnitude than the others. Therefore, CS3 and BHL3 are not the best performing estimators as the RMSEs place these at 5th and 12th in order of performance. Although the combination of SHL2 and CS2 is the second best in terms of RMSE, its correlation with the true spread is relatively low. SHL2's correlation coefficient is 0.63; this is acceptable in comparison with others. Also, its RMSE is the lowest amongst the models; therefore we rank this as the best performer. In Table 6, the average cross-sectional correlations between true and estimated spreads are reported across all currency pairs. CS3 and BHL3 exhibit the highest correlations, which are 0.959 and 0.955. SHL2's correlation is 0.81, performing slightly less well than CS3 and BHL3 in this instance but still at an adequate level. The average time series correlations between true and estimated spreads are reported

[1]Outliers are deemed to be those spreads which exceed the daily average by over 50 times.

across all currency pairs. BHL3 and AR exhibit the highest correlations, which are 0.6 and 0.53 respectively while with SHL2 the correlation is 0.03. The average time series correlations are much lower than those generated through cross-sectional analysis; this may be as a result of the time series being relatively short with its length being 9 months.

[Insert Tables 5 to 6]

4.4 Comparisons in equity markets TAQ data 2014

In this section, we use our chosen estimators to gauge spreads for the U.S equity market using the constituents of the S&P 1500 index as a sample. A snapshot of TAQ data, offers tick by tick pricing in 2014, which is used to calculate volume-weighted effective bid-ask spread and daily high and low prices. Table 7 reports the results of S&P 1500 (pooled case) stocks [m]. In the pooled case, the SHL2 significantly outperforms other estimators in terms of RMSE results. It must be noted that all estimators except for SHL2 significantly over or underestimate the spread. SHL2 displays the smallest estimated error and underestimates the spread by 19% on average, while the next best performing model (CS3) has an error of 282%. In contrast to a relatively poor performance with FX simulations, CS3 is ranked second out of all the estimators. The combination of SHL2 and BHL1 takes the third place. The pooled equity-month correlation coefficients of BHL3 and SHL2 are 0.82 and 0.75 respectively; this is significantly higher than the others and suggests a high correlation with the true spread. Table 8 reports the average ranking of the small, mid and large cap equity group cases. It is apparent that SHL2 is still the best choice for estimator while CS2 and the combination of SHL2 and BHL1 produce results that could offer a good alternative. Table 8 reports the equity-by-equity cross-sectional correlations for each of the estimators. BHL3 and SHL2's correlations are 0.84 and 0.77 respectively; these are significantly higher than the others reported through the testing. AR's correlation of 0.71 is the third strongest. Table 8 reports average time series correlations of the estimators. BHL3 and CS3's correlations are 0.34 and 0.29. SHL2's correlation is 0.06.

[m]The results of S&P 600 (small cap), S&P 400 (mid cap), S&P 500 (large cap) stocks are reported in online appendix.

[Insert Tables7 to 8 here]

5 Application of SHL2: NYSE 1926-2015

Moving beyond simulation, in this section, we demonstrate the application of SHL2 by using this estimator to gauge monthly average spreads for developed and emerging market stock exchanges. We find that the SHL2 acts as a good proxy for market liquidity as predictions of periods of intense uncertainty are often accompanied with low liquidity (high bid-ask spreads) levels in financial markets. In the samples we investigate, the data for the true spread is unavailable.

Figure 4 shows the monthly average estimated spread of all US stock markets including New York Stock Exchange (NYSE), American Stock Exchange (AMEX) and teh Nasdaq from 1926 to 2015; this is generated by SHL2 using daily CRSP data[n]. The monthly average estimated spread of each market are also shown separately. The spread was relatively large in the years before 1935. A further period of low liquidity can be observed from 1970 to 1992, which is mainly caused by the low liquidity in the Nasdaq. In 2008 at the nadir of the global financial crisis, the SHL2 estimator recorded considerable lower liquidity levels through increased spreads, than in the years surrounding the event.

[Insert figures 4 here]

The applications in this section suggest that SHL2 can act as an estimator that is sensitive enough to capture notable market events affecting transaction costs and as a consequence, liquidity levels. SHL2, as a spread estimator, can also be used as a liquidity measure in asset pricing models in a similar manner to that demonstrated in Corwin and Schultz (2012) and Abdi and Ranaldo (2017).

6 Conclusion

In this paper, we introduce two new low frequency bid-ask spread estimators which estimate the bid-ask spread using daily and two-day high and low prices. We show

[n] the monthly average estimated spreads for equities listed on the London, Hong Kong and Thai stock exchanges are shown in online appendix

that using similar input data, our estimators, in particular, the sophisticated version (SHL2), significantly outperforms both the latest and the popular models such as Abdi and Ranaldo (2017), Corwin and Schultz (2012) and Roll (1984) in terms of accuracy, efficiency, as well as cross-sectional and time series correlations.

We test the performance of estimators using comprehensive Monte Carlo simulation experiments under various 'signal to noise' ratios and different sampling frequencies. In addition, the cross-sectional properties of the estimators are also examined. Our estimators, BHL and SHL, appear to be unbiased throughout all tests carried out. By setting negative trials to zero which we label SHL2, we can obtain more efficient estimates; these are exhibited through lower standard errors. The results of simulation experiments suggest that our estimators outperform the AR, CS and Roll models. We demonstrate that SHL2 is the best single estimator in terms of accuracy and efficiency. We go further and test the performance of combinations of estimators against our own models and find that the AR, CS and Roll models also fail to match with ours in performance. The combinations of the estimators are useful as using these can address the problems associated with errors which often appear for individual models. The combination of basic and sophisticated high-low estimators (BHL1 and SHL2) perform well and offer a good alternative to using the single estimators, thereby avoiding the associated error risk.

We then move beyond simulation experiments to study the models using real world data for both foreign exchange and equity markets. We find that our SHL2 model outperforms all the others including the AR, CS and Roll models in terms of the root mean square error (RMSE). We verify this through running tests using trade and quote data for 23 currency pairs over 9 months and for equities listed on the S&P 500 throughout 2014. From these we show that SHL2's root mean square error (RMSE) is almost less than a half (even 20%) of the RMSE produced by other models. In terms of correlation, BHL3, AR, CS3 and SHL2 all performed well as estimators.

In general, our BHL and SHL are the best spread estimators. Researchers can choose the estimator according to their needs: BHL3 is good for cases where the high correlation is the only requirement and SHL2 can be used for other cases especially when accuracy and efficiency is of particular importance.

In order to demonstrate the effectiveness of our model (SHL2) through applica-

tions, we provide an illustration of how this can be applied. We generate the average monthly bid-ask spreads for the US, UK, HK and Thai equity markets. We show how the estimated spreads follow a pattern that is in line with our expectations in that the transaction costs increase sharply during crises periods.

Similar to the CS model, our estimators also obtain the estimates of daily mid-price diffusion at the same time as when the spreads are estimated. Because the spread and the diffusion are estimated together, a good spread estimator is also a good diffusion estimator. Thus, our sophisticated version model (SHL2) also offers the best diffusion estimates. As our estimators are not designed for a particular market structure, further research could test and apply our suggested estimators to the bonds, futures and option markets. In particular, these may be interesting for the over-the-counter markets where quote data can be difficult to obtain.

7 Appendix

7.1 Proof of Proposition 3.1

When the components of the spread do not include feedback trading, inventory control or asymmetric information, we can consider that the spread and its estimates, and thus the estimated errors, are either serially independent or fixed. If an estimate of the spread $\widetilde{SP}_i \in A$ corresponds to $Var_i = max(B)$, it equals the true spread i.e. $\widetilde{SP}_i = SP$.

Proof The variance of the conjectures of mid-price returns is:

$$\begin{aligned} Var_i &= Var\left[\Delta\widetilde{M}_t\right] \\ &= E\left\{\left[\Delta\widetilde{M}_t - E\left(\Delta\widetilde{M}_t\right)\right]^2\right\} \end{aligned} \tag{44}$$

We will assume that the expectation of the value of the conjectural mid-prices is zero. Thus, the equation above can be rewritten as:

$$\begin{aligned} &Var_i = Var\left(\Delta\widetilde{M}_t\right) \\ &= E\left(\Delta\widetilde{M}_t{}^2\right) \\ &= E\left[\left(\Delta M_t + \tfrac{1}{2}\Omega BS_t - \tfrac{1}{2}\Omega BS_{t-1}\right)^2\right] \\ &= E\left[\left(\Delta M_t + \tfrac{1}{2}\Omega BS_t - \tfrac{1}{2}\Omega BS_{t-1}\right)\left(\Delta M_t + \tfrac{1}{2}\Omega BS_t - \tfrac{1}{2}\Omega BS_{t-1}\right)\right] \\ &= E\left(\Delta M_t{}^2 + \tfrac{1}{2}\Omega BS_t\Delta M_t - \tfrac{1}{2}\Omega BS_{t-1}\Delta M_t\right) \\ &+E\left[\tfrac{1}{2}\Omega BS_t\Delta M_t + \tfrac{1}{4}(\Omega BS_t)^2 - \tfrac{1}{4}\Omega^2 BS_t BS_{t-1}\right] \\ &-E\left[\tfrac{1}{2}\Delta M_t\Omega BS_{t-1} + \tfrac{1}{4}\Omega^2 BS_t BS_{t-1} - \tfrac{1}{4}(\Omega BS_{t-1})^2\right] \end{aligned} \tag{45}$$

Where Ω denotes the conjectural error which represents the difference between the conjectural mid-price and the true mid-price, alternately expressed as the difference between the conjectural spread and its true value. Formally, Ω is given as:

$$\Omega = \Delta\widetilde{M}_t - \Delta M_t = \widetilde{SP}_i - SP \tag{46}$$

The assumptions of this proposition imply that BS is independent of ΔM at all observation points, therefore many of the terms in (45) such as $E(\Delta M_t\, BS_t)$ and $E(\Delta M_{t-1}\, BS_t)$ equate to zero. Formally, we have:

$$
\begin{aligned}
&E(\Delta M_t\, BS_t) = 0 \\
&E(\Delta M_{t-1}\, BS_t) = 0 \\
&E(\Delta M_t\, BS_{t-1}) = 0 \\
&E\left(BS_t\, BS_{t-1}\right) = 0
\end{aligned}
\tag{47}
$$

Furthermore, the variable BS is a binary variable (1 or -1), thus:

$$
E\left({BS_{t-1}}^2\right) = 1 \tag{48}
$$

Finally we obtain:

$$
\begin{aligned}
Var_i &= Var\left(\Delta\widetilde{M}_t\right) \\
&= E\left({\Delta M_t}^2 + \tfrac{1}{2}\,\Omega^2\right)
\end{aligned}
\tag{49}
$$

The final step of Equation (49) given above is the quadratic polynomial of the expectation of the error of the conjecture. For a given series, the first term $E(\Delta M_t^2)$ is a constant. We can surmise directly from this that when the error is zero (i.e. $\Omega = 0$), the second term $\frac{1}{2}\Omega^2$ is zero. Furthermore, when $\Omega = 0$, there is a global extreme for the right hand side polynomial in the final step, symmetrically, the left hand side of the equation $Var_i = Var\left(\Delta\widetilde{M}_t\right)$ is also at the extreme value. Formally this can be expressed as:

$$
\underset{\Omega}{arg\ max}\, Var\left(\Delta\widetilde{M}_t\right) = 0 \tag{50}
$$

When the conjectural error is zero, the conjectural spread becomes the true spread:

$$
\widetilde{SP}_i = SP + \Omega = SP \tag{51}
$$

Therefore the conjectural spread which maximises the covariance equals the true spread.

$$
\underset{\widetilde{SP}_i \in A}{arg\ max}\, Var\left(\Delta\widetilde{M}_t\right) = SP \tag{52}
$$

Q.E.D.

7.2 Proof of feedback bias

When feedback trading exists, we have:

$$
E(\Delta M_t\, BS_t) \neq 0 \tag{53}
$$

Substituting Equations (47), (48) and (53) and into Equation (45), we can obtain:

$$\begin{aligned}
&Var_i = Var\left[\Delta\widetilde{M}_t\right] \\
&= E\left[(\Delta M_t)^2 + \tfrac{1}{2}\Omega BS_t\Delta M_t - \tfrac{1}{2}\Omega BS_{t-1}\Delta M_t\right] \\
&+E\left[\tfrac{1}{2}\Omega BS_t\Delta M_t + \tfrac{1}{4}(\Omega BS_t)^2 - \tfrac{1}{4}\Omega^2 BS_t BS_{t-1}\right] \\
&-E\left[\Delta M_t\tfrac{1}{2}\Omega BS_{t-1} + \tfrac{1}{4}\Omega^2 BS_t BS_{t-1} - \tfrac{1}{4}(\Omega BS_{t-1})^2\right] \\
&= E\left(\Delta M_t{}^2 + \tfrac{1}{2}\Omega BS_t\Delta M_t\right) \\
&+E\left(\tfrac{1}{2}\Omega BS_t\Delta M_t + \tfrac{1}{2}\Omega^2\right) \\
&= E\left(\Delta M_t{}^2 + \Omega BS_t\Delta M_t + \tfrac{1}{2}\Omega^2\right)
\end{aligned} \tag{54}$$

Substituting $\Omega = \widetilde{SP} - SP$ into the equation above, we have:

$$\begin{aligned}
&Var_i = Var\left[\Delta\widetilde{M}_t\right] \\
&= E\left[(\Delta M_t)^2 + \Omega BS_t\Delta M_t + \tfrac{1}{2}\Omega^2\right] \\
&= E\left[(\Delta M_t)^2 + \left(\widetilde{SP} - SP\right)BS_t\Delta M_t + \tfrac{1}{2}\left(\widetilde{SP} - SP\right)^2\right]
\end{aligned} \tag{55}$$

Using first order conditioning of Equation (55), we obtain:

$$SP = \widetilde{SP} - E\left(BS_t\,\Delta\,M_t\right) \tag{56}$$

Equation above suggests that when there is feedback trading, variance version of the BL estimator overestimates the spread.

7.3 Proof of Equation (31)

For each day, we choose at random either the daily high or low prices to calculate the daily price change. Thus, the probability of picking daily high (or low) price is 50% and there are four cases for the daily price changes with an equal likelihood which are as follows.

$$\Delta\,M_{daily} = \begin{cases} H_t - H_{t-1} \;\; with\;\tfrac{1}{4}\;chance \\ H_t - L_{t-1} \;\; with\;\tfrac{1}{4}\;chance \\ L_t - H_{t-1} \;\; with\;\tfrac{1}{4}\;chance \\ L_t - L_{t-1} \;\; with\;\tfrac{1}{4}\;chance \end{cases} \tag{57}$$

Thus $BS_{daily} \cdot \Delta M_{daily}$ is given as follows:

$$BS_{daily} \cdot \Delta M_{daily} = \begin{cases} BS_{daily} \cdot (H_t - H_{t-1}) \ \ with \ \frac{1}{4} \ chance \\ BS_{daily} \cdot (H_t - L_{t-1}) \ \ with \ \frac{1}{4} \ chance \\ BS_{daily} \cdot (L_t - H_{t-1}) \ \ with \ \frac{1}{4} \ chance \\ BS_{daily} \cdot (L_t - L_{t-1}) \ \ with \ \frac{1}{4} \ chance \end{cases} \tag{58}$$

When daily high (or low) price is picked, trading direction is known (Equation 4). Formally, we have:

$$BS_{daily} \cdot \Delta M_{daily} = \begin{cases} [1 \cdot (H_t - H_{t-1})] \ \ with \ \frac{1}{4} \ chance \\ [1 \cdot (H_t - L_{t-1})] \ \ with \ \frac{1}{4} \ chance \\ [-1 \cdot (L_t - H_{t-1})] \ \ with \ \frac{1}{4} \ chance \\ [-1 \cdot (L_t - L_{t-1})] \ \ with \ \frac{1}{4} \ chance \end{cases} \tag{59}$$

Taking the expectation of $BS_{daily} \cdot \Delta M_{daily}$, we obtain:

$$\begin{aligned} &E\left(BS_{daily} \cdot \Delta M_{daily}\right) = \tfrac{1}{4} \cdot E\left(H_t - H_{t-1}\right) + \tfrac{1}{4} \cdot E\left(H_t - L_{t-1}\right) \\ &-\tfrac{1}{4} \cdot E\left(L_t - H_{t-1}\right) - \tfrac{1}{4} \cdot E\left(L_t - L_{t-1}\right) \\ &= \tfrac{1}{2} E\left(H_t - L_t\right) \end{aligned} \tag{60}$$

7.4 A brief introduction to the AR, Roll and CS estimators

Researchers generally opt to use the Roll estimator and models[o] derived from it because they are easy to program. The Roll estimator is given by the following equation.

$$SP = 2\sqrt{-cov\left(\Delta s_t, \ \Delta s_{t-1}\right)} \tag{61}$$

According to Corwin and Schultz (2012), the CS estimator appears to be the best of low-frequency estimators including the Lesmond et al. (1999) estimator. Furthermore, our proposed model in this paper shares the same intuition with the CS estimator, therefore, the CS estimator is picked to examine. Squaring both sides of Equation (7), we have,

$$\begin{aligned} &\left(Range^T_{daily}\right)^2 = \left(Range^M_{daily} + SP\right)^2 \\ &= \left(Range^M_{daily}\right)^2 + 2\, Range^M_{daily} \cdot SP + (SP)^2 \end{aligned} \tag{62}$$

[o]Related models include Glosten and Harris (1988), Choi et al. (1988), Stoll (1989), George et al. (1991), Huang and Stoll (1997), Hasbrouck (2004, 2009) and Chen et al. (2016)

Similarly, squaring both sides of Equation (8), we have:

$$\begin{aligned}\left(Range^{T}_{twoday}\right)^{2} &= \left(Range^{M}_{twoday} + SP\right)^{2} \\ &= \left(Range^{M}_{twoday}\right)^{2} + 2\, Range^{M}_{twoday} \cdot SP + (SP)^{2}\end{aligned} \tag{63}$$

Corwin and Schultz (2012) assume that

$$\begin{aligned}E\left(Range^{T}_{twoday}\right)^{2} &\approx E\left[\left(Range^{T}_{twoday}\right)^{2}\right] \\ E\left(Range^{T}_{daily}\right)^{2} &\approx E\left[\left(Range^{T}_{daily}\right)^{2}\right]\end{aligned} \tag{64}$$

One could solve the spread from the equation system and obtains:

$$SP = \frac{2\left(e^{\alpha} - 1\right)}{1 + e^{\alpha}} \tag{65}$$

where

$$\alpha = \frac{\sqrt{2\beta} - \sqrt{\beta}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}} \tag{66}$$

$$\beta = E\left\{\sum_{J=0}^{1}\left(Range^{T}_{daily,t+J}\right)^{2}\right\};\ \gamma = \left(Range^{T}_{twoday}\right)^{2} \tag{67}$$

When the spread is small, $SP \approx \alpha$. We may use Equation (66) to estimate the spread.

Abdi and Ranaldo (2017) model incorporates the CS model into the Roll estimator. Formally, it can be expressed as follows:

$$SP = 2\sqrt{\left(s_t - \eta_t\right)\left(s_t - \eta_{t+1}\right)} \tag{68}$$

where η is the mid-point of the high and low prices. Formally, it is given by:

$$\eta_t = \frac{H_t + L_t}{2} \tag{69}$$

7.5 Comparison with Quoted Spread

Chung and Zhang (2014) show that closing bid ask spreads are a very good proxy for the effective spread. Although it could be construed as a controversial approach to use this proxy as a benchmark, it is still interesting to compare the quoted and estimated spreads visually.

Figure 1 illustrates this comparison using the example of the estimates and the actual closing quoted spread in the form of the USD/JPY currency pair taken over a

50-month period. We collect the data from DataStream. In the figure, it is apparent that all estimators, except for SHL2, show negative estimates. CS2 and the combinations appear more volatile than those produced by SHL2. It is the combination of SHL2 and BHL1 that has the lowest average estimated error.

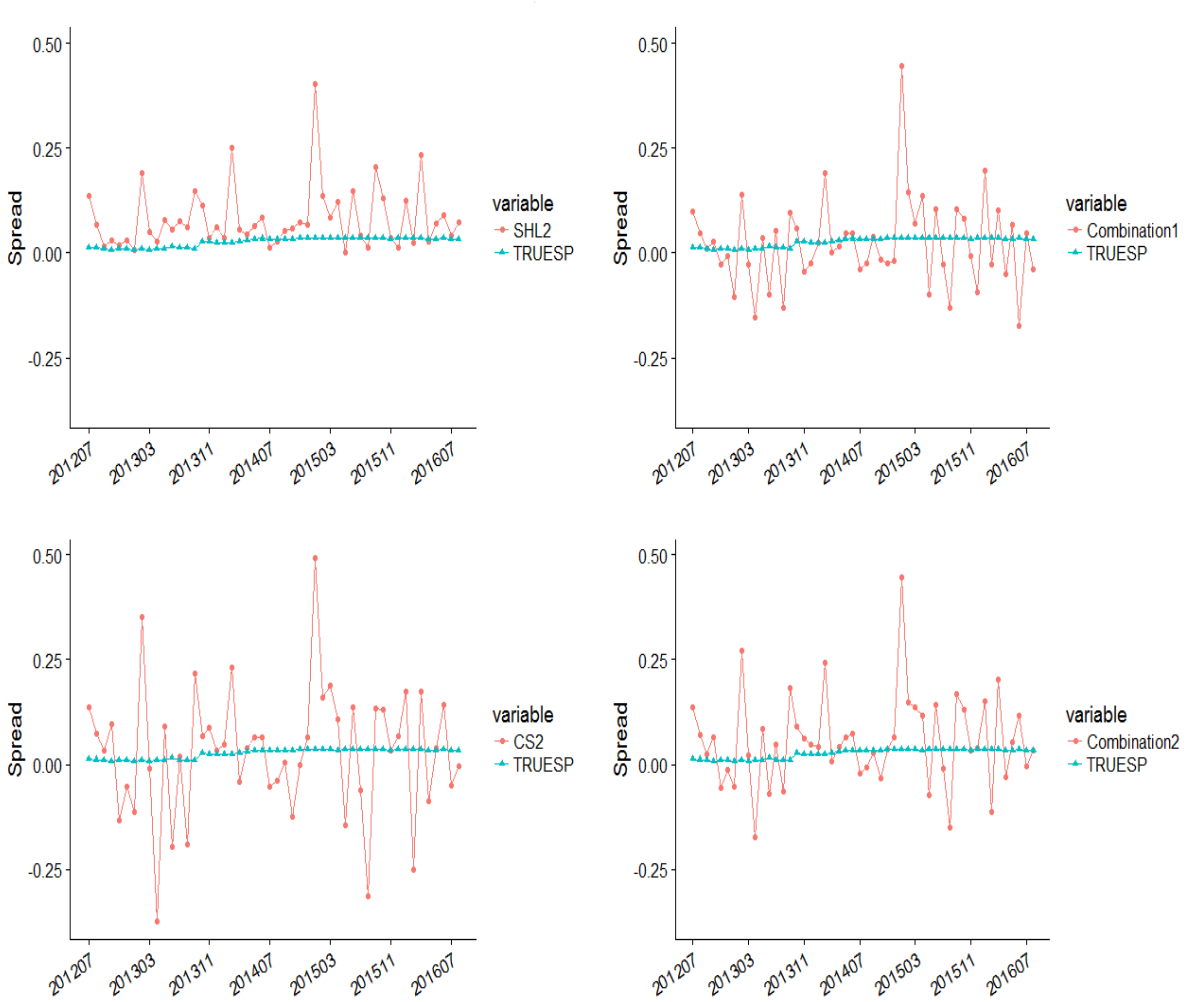

Figure 1: Monthly quoted and estimated spread, USD/JPY

The graphs above display estimates together with true values for a spread over a 50 month period from July 2012 to August 2016. The currency pair chosen for the illustration without losing generality is USD/JPY. True spreads (TRUESP) data are monthly average closing spreads taken from DataStream.

References

Abdi, F. and A. Ranaldo (2017). A Simple Estimation of Bid-Ask Spreads from Daily Close, High, and Low Prices. *University of St. Gallen, School of Finance Research Paper 2016/04*.

Bandi, F. M. and J. R. Russell (2006). Separating microstructure noise from volatility. *Journal of Financial Economics 79*(3), 655–692.

Banti, C., K. Phylaktis, and L. Sarno (2012). Global liquidity risk in the foreign exchange market. *Journal of International Money and Finance 31*(2), 267–291.

Bleaney, M. and Z. Li (2015). The performance of bid-ask spread estimators under less than ideal conditions Michael Bleaney and Zhiyong Li less than ideal conditions. *Studies in Economics and Finance 32*(1), 98–127.

Bleaney, M. and Z. Li (2016). A new spread estimator. *Review of Quantitative Finance and Accounting 47*(1), 179–211.

Chen, X., O. Linton, S. Schneeberger, and Y. Yi (2016). Simple Nonparametric Estimators for the Bid-Ask Spread in the Roll Model. *Cowles Foundation Discussion Paper No. 2033*, 1–59.

Choi, J. Y., D. Salandro, and K. Shastri (1988). On the Estimation of Bid-Ask Spreads: Theory and Evidence. *The Journal of Financial and Quantitative Analysis 23*(2), 219–230.

Chung, K. H. and H. Zhang (2014). A simple approximation of intraday spreads using daily data. *Journal of Financial Markets 17*(1), 94–120.

Corwin, S. A. and P. Schultz (2012). A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices. *Journal of Finance 67*(2), 719–760.

Fong, K. Y. L., C. W. Holden, and C. A. Trzcinka (2017). What Are The Best Liquidity Proxies For Global Research? *Review of Finance 6*, 1–22.

George, T. J., G. Kaul, and M. Nimalendran (1991). Estimation of the Bid-Ask Spread and its Components: A New Approach. *The Review of Financial Studies 4*(4), 623–656.

Glosten, L. R. and L. E. Harris (1988). Estimating the components of the bid/ask spread. *Journal of Financial Economics 21*(1), 123–142.

Goyenko, R. Y., C. W. Holden, and C. A. Trzcinka (2009). Do liquidity measures measure liquidity? *Journal of Financial Economics 92*(2), 153–181.

Harris, L. (1990). Statistical Properties of the Roll Serial Covariance Bid/Ask Spread Estimator. *The Journal of Finance 45*(2), 579–590.

Hasbrouck, J. (2004). Liquidity in the Futures Pits: Inferring Market Dynamics from Incomplete Data. *The Journal of Financial and Quantitative Analysis 39*(2), 305–326.

Hasbrouck, J. (2009). Trading Costs and Returns for U.S. Equities: Estimating Effective Costs from Daily Data. *The Journal of Finance 64*(3), 1445–1477.

Holden, C. W. (2009). New low-frequency spread measures. *Journal of Financial Markets 12*(4), 778–813.

Holden, C. W. and S. Jacobsen (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance 69*(4), 1747–1785.

Huang, R. D. and H. R. Stoll (1997). The Components of the Bid-Ask Spread: A General Approach. *The Review of Financial Studies 10*(4), 995–1034.

Karnaukh, N., A. Ranaldo, and P. Soderlind (2015). Understanding FX Liquidity. *Review of Financial Studies 28*(11), 3073–3108.

Lesmond, D. A., J. P. Ogden, and C. A. Trzcinka (1999). A New Estimate of Transaction Costs. *The Review of Financial Studies 12*(5), pp. 1113–1141.

Mancini, L., A. Ranaldo, and J. Wrampelmeyer (2013). Liquidity in the Foreign Exchange Market: Measurement, Commonality, and Risk Premiums. *The Journal of Finance 68*(5), 1805–1841.

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *The Journal of Business 53*(1), 61–65.

Roll, R. (1984). A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market. *The Journal of Finance 39*(4), 1127–1139.

Stoll, H. R. (1989). Inferring the Components of the Bid-Ask Spread: Theory and Empirical Tests. *The Journal of Finance 44*(1), 115–134.

Table 1: Definition of the Abbreviations in Tables

	Description	Calculation
SHL1	Sophisticated High and Low Model version 1	Calculates the spread using equation (34).
SHL2	Sophisticated High and Low Model version 2	Calculates the spread using equation (35).
BHL1	Basic High and Low Model version 1	Calculates the two-day interval spread using equation (14) and then calculates the monthly mean of the spread.
BHL2	Basic High and Low Model version 2	Calculates the average daily and two-day interval range for each month and then calculates the spread using equation (14).
BHL3	Basic High and Low Model version 3	Calculates the two-day interval spread using equation (14), and then calculates the monthly mean of the spread, letting all negative estimates equal to zero.
CS1	Corwin and Schultz (2012) model version 1	Calculates the two-day interval spread using equation (65) and then calculates the monthly mean of the spread.
CS2	Corwin and Schultz (2012) model version 2	Calculates the average daily and two-day interval range for each month and then calculates the spread using equation (65).
CS3	Corwin and Schultz (2012) model version 3	Calculates the two-day interval spread using equation (65) and then calculates the monthly mean of the spread. Letting all negative estimates equate to zero.
AR	Abdi and Ranaldo (2017) model	Calculates the spread using equation (68)
ROLL	Roll (1984) model	Calculates the spread using equation (61)
Combination 1	Combination of BHL1 and SHL2	(BHL1+SHL2)/2
Combination 2	Combination of SHL2 and CS2	(SHL12+CS2)/2
Combination 3	Combination of CS1 and BHL1	(CS1+BHL1)/2
Combination 4	Combination of BHL1 and BHL2	(BHL1+BHL2)/2

Table 2: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=30 (*0.001)	True spread/Midstd=0.158					True spread/Midstd=0.387				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)
SHL1	29.607	-0.61%	1.387	1.387	8	29.809	-0.603%	0.218	0.218	5
SHL2∗	52.870	74.94%	0.833	1.121	1	29.818	-0.588%	0.218	0.218	4
(negatives to be zero)										
BHL1⋆	29.567	-0.92%	1.375	1.375	6	29.820	-0.573%	0.218	0.218	3
(mean spreads)										
BHL2⋆	29.595	-0.74%	1.385	1.385	7	29.809	-0.607%	0.218	0.218	6
(mean parameters)										
BHL3⋆	128.947	330%	0.826	3.398	13	63.517	112%	0.142	1.126	14
(negatives to be zero)										
CS1ˆ	48.244	61.34%	1.249	1.391	9	36.619	22.091%	0.206	0.302	11
(mean spreads)										
CS2ˆ	30.054	0.83%	1.439	1.439	10	29.055	-3.116%	0.237	0.239	10
(mean parameters)										
CS3‡	100.202	234%	0.724	2.447	12	52.595	75.295%	0.135	0.765	13
(negatives to be zero)										
ROLL	95.550	416%	2.206	4.706	14	30.286	47.501%	0.582	0.751	12
AR‡	76.747	155%	0.750	1.724	11	35.313	17.669%	0.131	0.220	7
(negatives to be zero)										
Combination1	41.218	36.99%	1.067	1.129	2	29.819	-0.581%	0.217	0.217	1
(BHL1+SHL2)/2										
Combination2	41.462	37.88%	1.103	1.166	3	29.437	-1.856%	0.226	0.226	8
(SHL2+CS2)/2										
Combination3	38.905	30.19%	1.296	1.331	4	33.219	10.758%	0.211	0.237	9
(CS1+BHL1)/2										
Combination4	29.581	-0.83%	1.358	1.358	5	29.814	-0.590%	0.218	0.218	2
(BHL1+BHL2)/2										

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

†The column 'Ranking' reports the rankings of the estimators according to values produced in column RMSE.

This table reports the results of the time intervals of daily and four-hours respectively. *Midstd* represents the standard deviation of mid-price returns over the relevant interval. For each time interval, there are five panels which report the summary statistics and the results of the estimators respectively. Mean indicates the average of estimated spreads over 25000 replications. Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation We also report the rankings of the estimators according to RMSE.

Table 3: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=11.02 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0581						
Range from 1.00 to 20.00 (*0.001)	75000 months 20 observations per month						
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	10.461	-2.53%	5.612	5.612	7	0.113	10
SHL2∗	40.867	378%	4.006	5.506	5	**0.127**	2
(negatives to be zero)							
BHL1⋆	10.429	-3.79%	5.560	5.560	6	0.115	8
(mean spreads)							
BHL2⋆	10.438	-2.87%	5.618	5.618	8	0.114	9
(mean parameters)							
BHL3⋆	117.774	1343%	10.005	16.748	13	0.112	11
(negatives to be zero)							
CS1ˆ	30.265	242%	5.339	5.863	10	0.119	4
(mean spreads)							
CS2ˆ	11.820	14.74%	5.849	5.851	9	0.107	12
(mean parameters)							
CS3‡	88.586	979.10%	7.484	12.324	12	**0.136**	1
(negatives to be zero)							
ROLL	93.103	1759%	14.736	22.950	14	0.003	14
AR‡	74.796	825.24%	6.948	10.788	11	0.008	13
(negatives to be zero)							
Combination1	25.648	186%	4.286	4.673	1	0.124	3
(BHL1+SHL2)/2							
Combination2	26.343	195%	4.462	4.871	2	0.117	6
(SHL2+CS2)/2							
Combination3	20.347	119%	5.314	5.445	3	0.118	5
(CS1+BHL1)/2							
Combination4	10.434	-3.38%	5.499	5.499	4	0.116	7
(BHL1+BHL2)/2							

The standard divination of daily mid-price return is 0.1897

Mean indicates the average of estimated spreads over 75000 months. The true spread changes every month ranging from 0.002 to 0.02.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

†The column 'Ranking' reports the rankings of the estimators according to values produced in column RMSE.

This table reports the results of the time interval of daily. *Midstd* represents the standard deviation of mid-price returns over the relevant interval. For each time interval, there are five panels which report the summary statistics and the results of the estimators respectively. Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation.

Table 4: Simulation experiments: Average Ranking

	All cases according to simulations in sections 4.2.1 and 4.2.2 RMSE	Fixed spread cases according to simulations in section 4.2.1 RMSE	Cross-sectional cases according to simulations in section 4.2.2 RMES	Cross-sectional cases according to simulations in section 4.2.2 Correlation
SHL1	7.3	7.2	7.4	9.5
SHL2∗ (negatives to be zero)	2.4	2.1	3	6.67
BHL1 (mean spreads) ⋆	5.5	5.3	6	6
BHL2 (mean parameters) ⋆	6.9	6.6	7.6	7.33
BHL3 (negatives to be zero) ⋆	13.3	13.5	13	4.5
CS1ˆ(mean spreads)	9.4	9.6	9	6.67
CS2ˆ(mean parameters)	9.5	9.2	10	9.5
CS3‡ (negatives to be zero)	12.1	12.1	12	5.5
ROLL	13.6	13.4	14	13.67
AR‡ (negatives to be zero)	10.7	10.6	11	13.33
Combination1 (BHL1+SHL2)/2	1.3	1.5	1	4.5
Combination2 (SHL2+CS2)/2	2.7	3.1	2	6.5
Combination3 (CS1+BHL1)/2	5.7	6.5	4	5.83
Combination4 (BHL1+BHL2)/2	4.5	4.3	5	5.5

This table reports the average ranking of the estimators and the combinations in simulations experiments according to the columns of ranking in Tables 2 to 17.

Definition of the abbreviations are given by Table 1

Table 5: Hotspot 23 Currency pairs from 2015.12 to 2016.8

Effective Spread = 5.833 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	-62.309	-757%	9.450	12.090	10	-0.745	14
SHL2∗ (negatives to be zero)	19.666	230%	2.755	3.585	1	0.627	5
BHL1⋆ (mean spreads)	-55.091	-597%	8.655	10.495	7	-0.741	13
BHL2⋆ (mean parameters)	-59.170	-724%	9.780	12.150	11	-0.714	11
BHL3⋆ (negatives to be zero)	147.311	2146%	11.938	24.543	12	**0.949**	1
CS1ˆ (mean spreads)	8.293	213%	5.015	5.439	4	0.131	6
CS2ˆ (mean parameters)	-10.994	14%	6.123	6.109	5	-0.267	8
CS3‡ (negatives to be zero)	66.780	997%	5.085	11.189	8	**0.936**	2
ROLL	136.242	2121%	14.175	25.479	14	0.784	4
AR‡ (negatives to be zero)	168.857	2297%	10.747	25.351	13	0.914	3
Combination1 (BHL1+SHL2)/2	-17.712	-182%	4.916	5.229	3	-0.589	9
Combination2 (SHL2+CS2)/2	4.336	125%	4.050	4.230	2	0.048	7
Combination3 (CS1+BHL1)/2	-23.399	-194%	6.107	6.394	6	-0.601	10
Combination4 (BHL1+BHL2)/2	-57.130	-659%	9.111	11.226	9	-0.737	12

The results above refer to the testing of the following currency pairs: AUD/JPY, AUD/NZD, AUD/USD, EUR/AUD, EUR/CHF, EUR/GBP, EUR/JPY, EUR/NOK, EUR/PLN, EUR/SEK, EUR/USD, GBP/JPY, GBP/USD, NZD/USD, USD/CAD, USD/CHF, USD/JPY, USD/MXN, USD/NOK, USD/SEK, USD/SGD, USD/TRY, USD/ZAR. Tick by tick transaction and quoted data are used to generate the monthly effective spread.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 6: Currency-by-currency average correlation Hotspot from 2015.12 to 2016.8

	Time series correlation	Ranking (Time)	Cross-sectional correlation	Ranking (Cross)	Average Ranking (All)
SHL1	-0.405	14	-0.587	12	13.3
SHL2 (negatives to be zero) $*$	0.030	6	0.808	4	5
BHL1 (mean spreads) $\star$	-0.321	11	-0.668	14	12.7
BHL2 (mean parameters) $\star$	-0.356	13	-0.531	11	11.7
BHL3 (negatives to be zero) $\star$	**0.606**	1	**0.955**	2	1.3
CS1 (mean spreads) $\hat{}$	0.072	5	0.292	6	5.7
CS2 (mean parameters) $\hat{}$	-0.033	8	-0.189	8	8
CS3‡ (negatives to be zero)	0.470	3	**0.959**	1	2
ROLL	0.345	4	0.724	5	4.3
AR‡ (negatives to be zero)	**0.528**	2	0.946	3	2.7
Combination1 (BHL1+SHL2)/2	-0.258	10	-0.298	9	9.3
Combination2 (SHL2+CS2)/2	0.006	7	0.166	7	7
Combination3 (CS1+BHL1)/2	-0.193	9	-0.385	10	9.7
Combination4 (BHL1+BHL2)/2	-0.344	12	-0.611	13	12.3

This table reports the average time-series and cross-sectional correlations of the estimators and the combinations.

Currency pairs used in this table are listed in Table 5

Definition of the abbreviations are given by Table 1

Tick by tick transaction and quoted data are used to generate the monthly effective spread.

Highest two correlation coefficients are made bold.

Ranking (Time) and Ranking (Cross) represent the rankings of time series, cross-sectional correlation respectively. Average Ranking (all) is the average ranking of correlation in Tables 5 and 6.

Table 7: TAQ (Effective Spread) S&P 1500 2014.01-2014.12

Effective Spread = 128.0 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	-264	-571%	6.245	8.461	11	-0.043	12
SHL2∗ (negatives to be zero)	69	-19.42%	0.908	0.929	1	**0.745**	2
BHL1⋆ (mean spreads)	-277	-592%	6.231	8.597	12	-0.044	14
BHL2⋆ (mean parameters)	-257	-561%	6.249	8.396	9	-0.043	11
BHL3⋆ (negatives to be zero)	359	452%	4.025	6.051	7	**0.823**	1
CS1ˆ (mean spreads)	-46	-297%	3.785	4.811	4	0.301	6
CS2ˆ (mean parameters)	-237	-606%	7.197	9.409	13	0.024	9
CS3‡ (negatives to be zero)	221	282%	2.776	3.958	2	0.519	4
ROLL	429	964%	9.737	13.702	14	0.252	7
AR ‡ (negatives to be zero)	307	413%	4.250	5.929	6	0.680	3
Combination1 (BHL1+SHL2)/2	-104	-305%	3.218	4.434	3	0.070	8
Combination2 (SHL2+CS2)/2	-84	-312%	3.730	4.865	5	0.318	5
Combination3 (CS1+BHL1)/2	-161	-446%	4.925	6.647	8	-0.002	10
Combination4 (BHL1+BHL2)/2	-267	-576%	6.191	8.459	10	-0.044	13

Tick by tick effective data are used to generate the monthly effective spread.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 8: Stock-by-stock average correlation S&P 1500 2014.01-2014.12

	Time series correlation	Ranking	Cross-sectional correlation	Ranking	All cases (RMSE)	All cases (Corr)
SHL1	-0.124	13	-0.037	11	10.3	12
SHL2 (negatives to be zero) $*$	0.056	5	**0.765**	2	1	3
BHL1 (mean spreads) $\star$	-0.115	11	-0.049	13	12	12.7
BHL2 (mean parameters) $\star$	-0.127	14	-0.047	12	9	12.3
BHL3 (negatives to be zero) $\star$	**0.334**	1	**0.844**	1	6.7	1
CS1 (mean spreads) $\hat{}$	-0.015	6	0.323	6	4	6
CS2 (mean parameters) $\hat{}$	-0.103	10	0.034	9	13	9.3
CS3‡ (negatives to be zero)	**0.294**	2	0.543	4	2	3.3
AR‡ (negatives to be zero)	0.102	4	0.36	5	6	3
ROLL	0.292	3	0.706	3	14	5.3
Combination1 (BHL1+SHL2)/2	-0.093	9	0.141	8	3	8.3
Combination2 (SHL2+CS2)/2	-0.083	7	0.318	7	5.3	6.3
Combination3 (CS1+BHL1)/2	-0.084	8	0.025	10	8	9.3
Combination4 (BHL1+BHL2)/2	-0.123	12	-0.049	13	10.7	12.7

This table reports the average time-series and cross-sectional correlation of the estimators and the combinations.

Tick by tick quoted data are used to generate the monthly quoted spread.

This table also reports the average ranking of the estimators and the combinations according to the columns of ranking (RMSE) in Tables 19 to 21 and ranking (Corr) in Tables 7 and 8 respectively.

Definition of the abbreviations are given by Table 1

Highest two correlation coefficients are made bold.

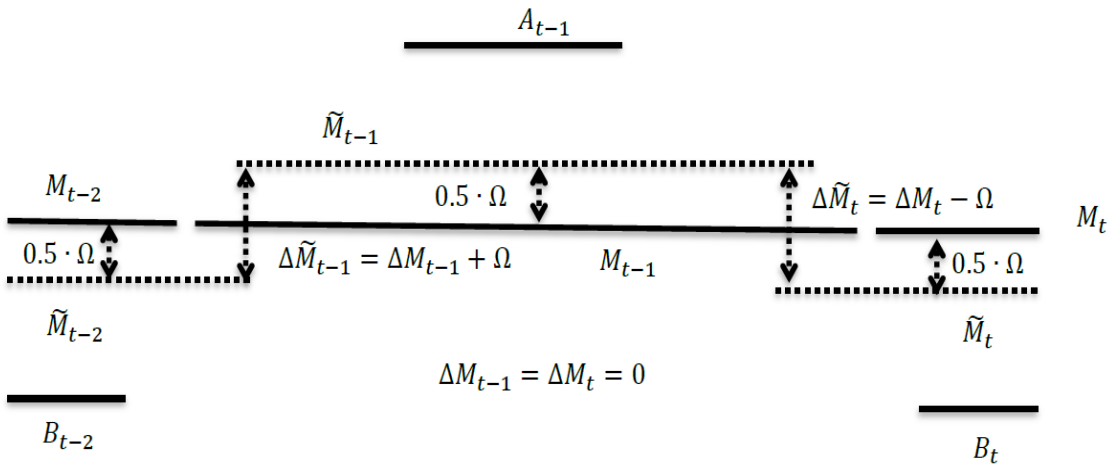

Figure 2: The Conjecture of the Spread

Source Bleaney and Li (2016)

Figure 2 outlines the reasoning underpinning this proposition where for the purposes of economy we hold that the mid-price is fixed. The conjectural spread ($\widetilde{SP}_i$) is less than the true spread. This allows us to estimate the conjectural mid-price $\widetilde{M}$; this is represented by the dotted line in Figure 2, and the true mid price and transaction price are both represented by unbroken lines. Also in Figure 2, A and B denote observed ask and bid prices, whereas M is the unobserved true mid-price. In addition, Δ is taken to be the first-order difference operator and Ω denotes the conjectural error.

At any one point we can only observe one price, either the bid or ask. In Figure 2, three periods are displayed. In the period labelled $t-2$, the bid price is recorded and in period labelled $t-1$, the ask price is observed. In period $t-2$, the conjectural spread is lower than the true spread and the conjectural mid-price error is -0.5Ω , which is less than the true value. In period $t-1$, the conjectural mid-price error is 0.5Ω, therefore this is greater than the true one. In the intervening period between $t-2$ and $t-1$, the direction of the trade shifts from sell to buy, and because of the conjectural error, we overestimate the mid-price return, formally we express this as:

$$\Delta\widetilde{M}_{t-1} = \Delta M_{t-1} + \Omega = \Omega$$

In Figure 2, the hypothetical example shows that the variance of mid-price returns equates to zero because returns remain fixed. However the variance of conjectured mid-price returns is greater than zero. The reason for this is that in the case where the spread is underestimated, the conjectured mid-price fluctuates more than its true counterparts.

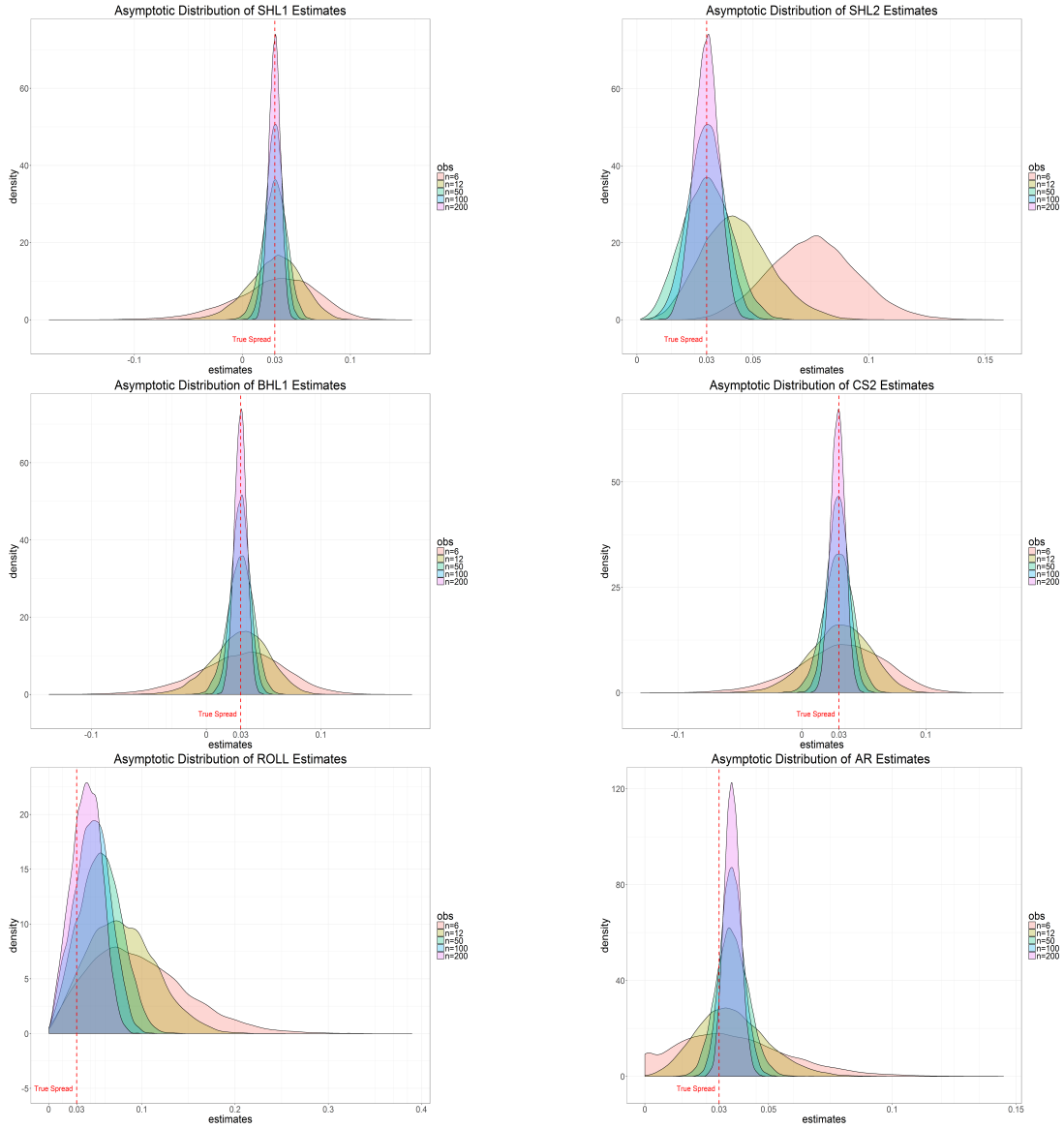

Figure 3: Asymptotic Distribution

This figure shows the distributions of the estimates in various groups. The simulation data used in the right panel (four-hour case) in Table (2) is selected to construct the asymptotic distributions of the estimators. In existing tests, for the results are given in table 2, there are 120 observations in a group per month. To demonstrate the asymptotic characteristics of the results produced through the simulation we consider five cases where there are 6, 12, 50, 100 and 200 observations for each group. We then generate estimated spreads according to each case (note there are 15000 groups in each).

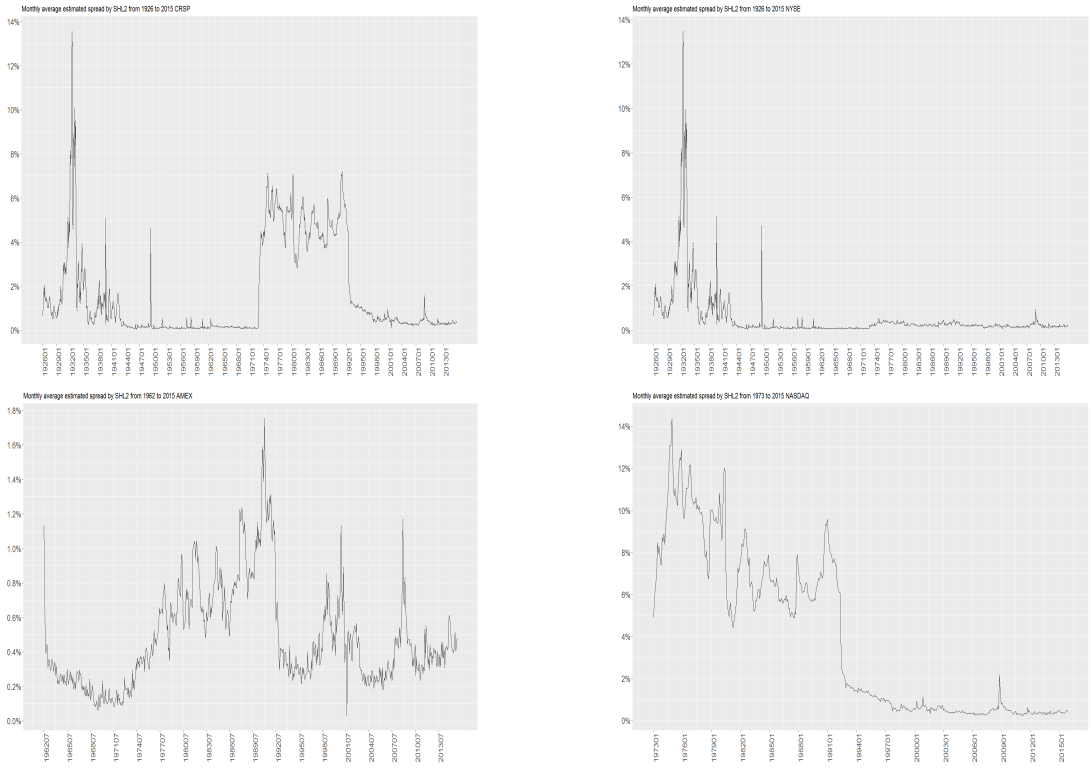

Figure 4: Monthly average estimated spread by SHL2 from 1926 to 2015, CRSP

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the New York Stock Exchange American Stock Exchange and Nasdaq on a monthly basis from January 1926 to December 2015. The figure plots the monthly equally weighted average spread of all stocks with each recording at least 16 daily spread observations within the month. All data is taken from CRSP.

Online appendix: Additional tables and figures

Autocorrelated Mid-Price Returns

In this section, we run simulation experiments when the mid-price returns are autocorrelated. The settings are similar to those in section 4.2.1 with the exception of a correlation of the mid-price of returns. In our estimation we only consider the case of negative autocorrelation; the positive one is symmetrically reflected. Formally, the settings are given as follows.

$$
\begin{aligned}
s_t &= +\tfrac{SP}{2} \cdot BS_t \\
BS_t &\sim B(1, 0.5) \\
\Delta M_t &= -0.33\, \Delta M_{t-1} + \epsilon_t \\
\epsilon_t &\sim N(0, 0.05)\ (one - minute) \\
SP &= 0.030
\end{aligned}
$$

Where ϵ is a random shock. The results for this simulation are presented in Table 9. As we can see that our estimators are still the best performing estimators. In general, the estimators are not influenced significantly by the autocorrelated spread. In comparison with those in Table (2) in the paper, the RMSE is smaller than those in (2), meaning that under the circumstance of autocorrelation the estimators perform marginally better than in the random walk case.

Table 9: Simulation experiments: Autocorrelated Mid-Price Returns

Mean true spread=30 (*0.001)	Daily (MidStd=179.9*0.001) Truespread/Midstd=0.1668				
	15000 months 20 observations per month				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)
SHL1	29.853	0.029%	1.028	1.028	3
SHL2∗	44.799	48%	0.658	0.816	1
(negatives to be zero)					
BHL1⋆	29.817	-0.242%	1.021	1.020	2
(mean spreads)					
BHL2⋆	29.860	0.009%	1.028	1.028	4
(mean parameters)					
BHL3⋆	101.525	238%	0.631	2.465	9
(negatives to be zero)					
CS1ˆ	43.453	45.17%	0.934	1.037	5
(mean spreads)					
CS2ˆ	30.130	0.818%	1.070	1.070	6
(mean parameters)					
CS3‡	80.099	167%	0.558	1.759	8
(negatives to be zero)					
ROLL	73.723	295%	1.672	3.391	10
AR‡	58.902	95.9%	0.568	1.115	7
(negatives to be zero)					

Mean indicates the average of estimated spreads over 15000 months. The true spread is the same across different months.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (3).

Additional tables

The following tables will not be reported in the main body for brevity. They will be available online.

Table 10: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread= 1 (*0.001)	True spread/Midstd=0.00527					True spread/Midstd=0.0129				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)†	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)†
SHL1	0.953	17.76%	41.177	41.176	8	0.899	-8.763%	6.513	6.514	7
SHL2∗	35.348	3391%	20.215	39.485	3	4.512	341%	3.609	4.968	1
(negatives to be zero)										
BHL1⋆	0.756	-6.63%	40.958	40.958	6	0.906	-7.918%	6.495	6.496	5
(mean spreads)										
BHL2⋆	0.793	0.97%	41.047	41.046	7	0.900	-8.702%	6.510	6.511	6
(mean parameters)										
BHL3⋆	112.316	11122%	23.817	113.742	13	46.195	4519%	3.775	45.349	14
(negatives to be zero)										
CS1ˆ	21.221	2039%	37.204	42.426	9	9.155	817%	6.100	10.195	10
(mean spreads)										
CS2ˆ	2.643	185%	43.125	43.164	10	0.999	1.231%	7.139	7.139	8
(mean parameters)										
CS3‡	82.901	8180%	19.591	84.112	12	34.266	3326%	3.221	33.415	12
(negatives to be zero)										
ROLL	93.496	15269%	64.917	165.918	14	20.969	3758%	15.947	40.823	13
AR‡	74.679	7350%	22.047	76.738	11	30.438	2943%	3.503	29.635	11
(negatives to be zero)										
Combination1	18.052	1693%	29.323	33.859	1	2.709	167%	4.972	5.244	2
(BHL1+SHL2)/2										
Combination2	18.996	1789%	30.546	35.397	2	2.756	171%	5.247	5.520	3
(SHL2+CS2)/2										
Combination3	10.988	1016%	38.554	39.869	4	5.030	404%	6.262	7.454	9
(CS1+BHL1)/2										
Combination4	0.775	-2.55%	40.307	40.306	5	0.903	-8.297%	6.485	6.485	4
(BHL1+BHL2)/2										

The standard divination of daily mid-price return is 0.1897. The true spread is fixed.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

†The column 'Ranking' reports the rankings of the estimators according to values produced in column RMSE.

This table reports the results of the time intervals of daily and four-hours respectively. *Midstd* represents the standard deviation of mid-price returns over the relevant interval. For each time interval, there are five panels which report the summary statistics and the results of the estimators respectively. Mean indicates the average of estimated spreads over 25000 replications. Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation We also report the rankings of the estimators according to RMSE.

Table 11: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=2 (*0.001)	True spread/Midstd=0.0105					True spread/Midstd=0.0258				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)
SHL1	2.181	20.22%	20.577	20.578	8	1.909	-3.972%	3.272	3.272	7
SHL2∗	36.002	1680%	10.134	19.620	3	5.096	150%	1.951	2.459	1
(negatives to be zero)										
BHL1⋆	1.985	8.91%	20.431	20.431	6	1.909	-4.022%	3.263	3.263	5
(mean spreads)										
BHL2⋆	2.169	19.27%	20.546	20.547	7	1.911	-3.887%	3.270	3.270	6
(mean parameters)										
BHL3⋆	113.081	5550%	11.873	56.758	13	46.773	2238%	1.912	22.466	14
(negatives to be zero)										
CS1ˆ	22.457	1032%	18.541	21.217	9	10.116	406%	3.069	5.092	10
(mean spreads)										
CS2ˆ	3.767	97.52%	21.513	21.535	10	1.967	-1.012%	3.593	3.593	8
(mean parameters)										
CS3‡	83.778	4085%	9.757	41.997	12	34.880	1644%	1.635	16.518	12
(negatives to be zero)										
ROLL	93.022	7581%	32.519	82.487	14	21.411	1839%	8.006	20.059	13
AR‡	74.919	3637%	10.969	37.985	11	30.540	1426%	1.756	14.372	11
(negatives to be zero)										
Combination1	18.993	845%	14.672	16.929	1	3.503	72.871%	2.571	2.672	2
(BHL1+SHL2)/2										
Combination2	19.884	889%	15.277	17.674	2	3.531	74.460%	2.712	2.812	3
(SHL2+CS2)/2										
Combination3	12.221	520%	19.240	19.930	4	6.012	201%	3.148	3.736	9
(CS1+BHL1)/2										
Combination4	2.077	14.08%	20.155	20.155	5	1.910	-3.949%	3.257	3.258	4
(BHL1+BHL2)/2										

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (10).

Table 12: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=6 (*0.001)	True spread/Midstd=0.0316					True spread/Midstd=0.0775				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)
SHL1	5.896	1.82%	6.900	6.900	8	5.818	-2.781%	1.080	1.081	7
SHL2∗	38.141	529%	3.523	6.354	3	7.634	25.770%	0.792	0.833	1
(negatives to be zero)										
BHL1⋆	5.805	-0.41%	6.831	6.831	6	5.821	-2.709%	1.080	1.080	5
(mean spreads)										
BHL2⋆	5.925	1.97%	6.896	6.895	7	5.817	-2.782%	1.080	1.080	6
(mean parameters)										
BHL3⋆	115.153	1818%	3.966	18.608	13	48.999	717%	0.646	7.195	14
(negatives to be zero)										
CS1ˆ	25.936	335%	6.219	7.063	9	13.797	130%	1.014	1.650	10
(mean spreads)										
CS2ˆ	7.377	25.97%	7.223	7.227	10	5.714	-4.482%	1.184	1.184	8
(mean parameters)										
CS3‡	85.942	1331%	3.315	13.713	12	37.168	519%	0.561	5.224	12
(negatives to be zero)										
ROLL	93.559	2455%	10.843	26.840	14	21.573	546%	2.646	6.071	13
AR‡	74.826	1144%	3.659	12.013	11	30.687	411%	0.586	4.154	11
(negatives to be zero)										
Combination1	21.973	264%	4.975	5.633	1	6.728	11.535%	0.927	0.935	2
(BHL1+SHL2)/2										
Combination2	22.759	277%	5.193	5.888	2	6.674	10.668%	0.971	0.977	3
(SHL2+CS2)/2										
Combination3	15.870	167%	6.443	6.656	4	9.809	63.730%	1.041	1.221	9
(CS1+BHL1)/2										
Combination4	5.865	0.82%	6.750	6.750	5	5.819	-2.740%	1.076	1.077	4
(BHL1+BHL2)/2										

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (10).

Table 13: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=10 (*0.001)	True spread/Midstd=0.0527					True spread/Midstd=0.129				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)
SHL1	9.884	0.82%	4.143	4.143	8	9.728	-2.641%	0.657	0.657	6
SHL2∗	40.456	301%	2.182	3.714	3	10.699	6.274%	0.552	0.556	1
(negatives to be zero)										
BHL1⋆	9.829	-0.04%	4.097	4.097	6	9.730	-2.629%	0.656	0.657	5
(mean spreads)										
BHL2⋆	9.833	0.15%	4.133	4.133	7	9.729	-2.626%	0.657	0.658	7
(mean parameters)										
BHL3⋆	117.638	1076%	2.385	11.017	13	51.223	412%	0.393	4.141	14
(negatives to be zero)										
CS1ˆ	29.716	199%	3.725	4.222	9	17.474	74.802%	0.617	0.970	10
(mean spreads)										
CS2ˆ	11.212	14.09%	4.313	4.315	10	9.454	-5.372%	0.718	0.720	8
(mean parameters)										
CS3‡	88.401	783%	2.014	8.085	12	39.539	295%	0.349	2.974	12
(negatives to be zero)										
ROLL	93.914	1433%	6.524	15.746	14	21.875	293%	1.611	3.343	13
AR‡	75.142	650%	2.200	6.860	11	30.997	210%	0.355	2.128	11
(negatives to be zero)										
Combination1	25.142	150%	3.018	3.371	1	10.214	1.822%	0.601	0.601	2
(BHL1+SHL2)/2										
Combination2	25.834	157%	3.140	3.512	2	10.076	0.455%	0.626	0.626	3
(SHL2+CS2)/2										
Combination3	19.772	99.33%	3.862	3.987	4	13.602	36.083%	0.633	0.729	9
(CS1+BHL1)/2										
Combination4	9.831	0.06%	4.048	4.048	5	9.729	-2.628%	0.655	0.655	4
(BHL1+BHL2)/2										

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

†The column 'Ranking' reports the rankings of the estimators according to values produced in column RMSE.

This table reports the results of the time intervals of daily and four-hours respectively. *Midstd* represents the standard deviation of mid-price returns over the relevant interval. For each time interval, there are five panels which report the summary statistics and the results of the estimators respectively. Mean indicates the average of estimated spreads over 25000 replications. Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation We also report the rankings of the estimators according to RMSE.

Table 14: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=3.57 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0188						
Range from 1.00 to 5.13 (*0.001)	15000 months 20 observations per month						
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	3.403	3.51%	12.511	12.511	7	0.010	12
SHL2∗	36.725	983%	6.814	11.957	3	0.010	11
(negatives to be zero)							
BHL1⋆	3.253	-3.09%	12.375	12.375	6	0.013	8
(mean spreads)							
BHL2⋆	3.494	4.31%	12.543	12.543	8	0.015	5
(mean parameters)							
BHL3⋆	113.768	3299.10%	11.395	34.903	13	**0.019**	1
(negatives to be zero)							
CS1ˆ	23.663	605.75%	11.388	12.899	9	0.017	3
(mean spreads)							
CS2ˆ	4.996	47.54%	13.054	13.062	10	**0.018**	2
(mean parameters)							
CS3‡	84.477	2423%	8.860	25.801	12	0.012	10
(negatives to be zero)							
ROLL	93.085	4448%	22.687	49.929	14	0.003	13
AR‡	74.800	2133%	8.872	23.101	11	-0.005	14
(negatives to be zero)							
Combination1	19.989	489%	9.053	10.291	1	0.013	9
(BHL1+SHL2)/2							
Combination2	20.861	515%	9.429	10.742	2	0.016	4
(SHL2+CS2)/2							
Combination3	13.458	301%	11.701	12.082	4	0.015	6
(CS1+BHL1)/2							
Combination4	3.373	0.36%	12.252	12.251	5	0.015	7
(BHL1+BHL2)/2							

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.002 to 0.00513.
Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (3).

Table 15: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=6.71 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0354						
Range from 5.13 to 8.29 (*0.001)	15000 months 20 observations per month						
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	6.533	0.62%	6.311	6.310	8	0.022	8
SHL2∗	38.490	476%	3.296	5.786	3	0.020	12
(negatives to be zero)							
BHL1⋆	6.521	-1.06%	6.268	6.268	6	**0.026**	1
(mean spreads)							
BHL2⋆	6.458	-0.96%	6.294	6.294	7	0.024	6
(mean parameters)							
BHL3⋆	115.408	1646%	4.232	16.993	13	0.024	7
(negatives to zero)							
CS1ˆ	26.490	302%	5.709	6.457	9	0.025	5
(mean spreads)							
CS2ˆ	7.966	21.36%	6.610	6.613	10	0.023	9
(mean parameters)							
CS3‡	86.192	1203%	3.458	12.518	12	0.022	11
(negatives to zero)							
ROLL	93.129	2224%	10.244	24.486	14	0.001	14
AR‡	74.634	1028%	3.641	10.905	11	0.005	13
(negatives to be zero)							
Combination1	22.505	237%	4.573	5.152	1	0.025	4
(BHL1+SHL2)/2							
Combination2	23.228	248%	4.765	5.374	2	0.022	10
(SHL2+CS2)/2							
Combination3	16.505	150%	5.908	6.095	4	**0.026**	3
(CS1+BHL1)/2							
Combination4	6.489	-0.96%	6.177	6.177	5	**0.026**	2
(BHL1+BHL2)/2							

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.00513 to 0.00829.
Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (3).

Table 16: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=9.86 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0520						
Range from 8.29 to 11.44 (*0.001)	15000 months 20 observations per month						
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	8.527	-11.13%	4.249	4.250	7	0.019	9
SHL2∗	39.773	302%	2.218	3.744	3	**0.021**	1
(negatives to be zero)							
BHL1⋆	8.584	-10.81%	4.222	4.223	6	0.018	7
(mean spreads)							
BHL2⋆	8.323	-13.21%	4.251	4.253	8	0.018	11
(mean parameters)							
BHL3⋆	116.896	1093%	2.644	11.244	13	0.020	3
(negatives to be zero)							
CS1ˆ	28.435	192%	3.836	4.289	9	0.018	10
(mean spreads)							
CS2ˆ	9.763	1.38%	4.445	4.445	10	0.016	12
(mean parameters)							
CS3‡	87.575	793%	2.191	8.230	12	**0.021**	2
(negatives to be zero)							
ROLL	92.115	1447%	6.774	15.979	14	-0.002	14
AR‡	74.396	658%	2.345	6.987	11	0.004	13
(negatives to be zero)							
Combination1	24.178	145%	3.090	3.415	1	0.020	4
(BHL1+SHL2)/2							
Combination2	24.768	152%	3.214	3.554	2	0.018	5
(SHL2+CS2)/2							
Combination3	18.509	90.50%	3.978	4.079	4	0.018	8
(CS1+BHL1)/2							
Combination4	8.453	-12.01%	4.169	4.171	5	0.018	6
(BHL1+BHL2)/2							

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.00829 to 0.0114.
Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (3).

Table 17: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=12.99 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0685						
Range from 11.45 to 14.55 (*0.001)	15000 months 20 observations per month						
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	12.128	-4.66%	3.218	3.218	8	0.020	10
SHL2∗	41.823	220%	1.730	2.798	3	0.025	3
(negatives to be zero)							
BHL1⋆	12.094	-5.31%	3.198	3.198	6	0.018	11
(mean spreads)							
BHL2⋆	12.081	-5.39%	3.217	3.217	7	0.020	7
(mean parameters)							
BHL3⋆	118.783	818%	1.939	8.403	13	**0.028**	1
(negatives to be zero)							
CS1ˆ	31.829	147%	2.907	3.258	9	0.020	6
(mean spreads)							
CS2ˆ	13.532	5.84%	3.363	3.364	10	0.018	12
(mean parameters)							
CS3‡	89.523	591%	1.642	6.137	12	**0.027**	2
(negatives to be zero)							
ROLL	92.732	1086%	5.084	11.992	14	0.010	14
AR‡	74.802	477%	1.745	5.080	11	0.012	13
(negatives to be zero)							
Combination1	26.958	107%	2.370	2.601	1	0.022	4
(BHL1+SHL2)/2							
Combination2	27.678	113%	2.463	2.709	2	0.021	5
(SHL2+CS2)/2							
Combination3	21.961	70.83%	3.013	3.095	4	0.020	8
(CS1+BHL1)/2							
Combination4	12.087	-5.39%	3.156	3.156	5	0.020	9
(BHL1+BHL2)/2							

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.0114 to 0.0146.
Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (3).

Table 18: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=16.10 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0849						
Range from 14.55 to 17.65 (*0.001)	15000 months 20 observations per month						
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	15.941	0.68%	2.590	2.590	7	0.010	8
SHL2∗	44.044	172%	1.425	2.232	3	0.008	11
(negatives to be zero)							
BHL1⋆	15.980	0.71%	2.577	2.577	6	**0.013**	1
(mean spreads)							
BHL2⋆	15.981	0.82%	2.596	2.596	8	0.011	6
(mean parameters)							
BHL3⋆	120.740	652%	1.559	6.699	13	**0.012**	4
(negatives to be zero)							
CS1ˆ	35.399	122%	2.346	2.642	9	0.009	12
(mean spreads)							
CS2ˆ	17.234	8.60%	2.718	2.720	10	0.009	10
(mean parameters)							
CS3‡	91.672	470%	1.316	4.885	12	0.011	7
(negatives to be zero)							
ROLL	93.978	862%	4.100	9.543	14	-0.010	13
AR‡	74.862	365%	1.409	3.915	11	-0.010	14
(negatives to be zero)							
Combination1	30.012	86.30%	1.928	2.112	1	**0.012**	3
(BHL1+SHL2)/2							
Combination2	30.639	90.25%	2.008	2.201	2	0.009	9
(SHL2+CS2)/2							
Combination3	25.689	61.12%	2.431	2.507	4	0.011	5
(CS1+BHL1)/2							
Combination4	15.981	0.78%	2.546	2.546	5	**0.012**	2
(BHL1+BHL2)/2							

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.0146 to 0.0177.
Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (3).

Table 19: TAQ (Effective Spread) S&P 500 2014.01-2014.12

Effective Spread = 202.3 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	-466	-1019%	8.063	12.994	11	-0.240	10
SHL2∗ (negatives to be zero)	55	6.86%	1.231	1.233	1	0.541	4
BHL1⋆ (mean spreads)	-471	-1043%	7.938	13.107	12	-0.262	13
BHL2⋆ (mean parameters)	-455	-1005%	8.113	12.913	9	-0.256	11
BHL3⋆ (negatives to be zero)	419	799%	4.342	9.094	6	**0.852**	1
CS1ˆ (mean spreads)	-106	-516%	5.145	7.288	4	0.218	6
CS2ˆ (mean parameters)	-370	-1076%	9.436	14.310	13	-0.273	14
CS3‡ (negatives to be zero)	242	515%	3.008	5.966	2	0.626	3
ROLL	575	1776%	11.660	21.247	14	0.352	5
AR‡ (negatives to be zero)	403	792%	4.646	9.182	7	**0.820**	2
Combination1 (BHL1+SHL2)/2	-208	-517%	4.224	6.677	3	-0.226	8
Combination2 (SHL2+CS2)/2	-157	-534%	5.004	7.317	5	-0.091	7
Combination3 (CS1+BHL1)/2	-289	-783%	6.391	10.106	8	-0.239	9
Combination4 (BHL1+BHL2)/2	-463	-1024%	7.937	12.952	10	-0.259	12

Tick by tick effective data are used to generate the monthly effective spread.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 20: TAQ (Effective Spread) S&P 400 2014.01-2014.12

Effective Spread = 365 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	-208	-471%	4.729	6.673	10	0.238	13
SHL2∗ (negatives to be zero)	78	-22.18%	0.825	0.854	1	0.826	3
BHL1⋆ (mean spreads)	-226	-496%	4.735	6.853	12	0.271	10
BHL2⋆ (mean parameters)	-197	-458%	4.746	6.596	9	0.265	12
BHL3⋆ (negatives to be zero)	358	382%	2.953	4.827	7	**0.910**	1
CS1ˆ (mean spreads)	-45	-253%	2.993	3.920	4	0.361	8
CS2ˆ (mean parameters)	-222	-504%	5.695	7.603	13	0.158	14
CS3‡ (negatives to be zero)	218	237%	2.074	3.151	2	0.541	6
ROLL	388	797%	7.313	10.815	14	0.417	7
AR‡ (negatives to be zero)	290	329%	2.980	4.441	6	**0.857**	2
Combination1 (BHL1+SHL2)/2	-74	-258%	2.502	3.596	3	0.543	5
Combination2 (SHL2+CS2)/2	-72	-263%	3.007	3.993	5	0.561	4
Combination3 (CS1+BHL1)/2	-135	-376%	3.805	5.351	8	0.322	9
Combination4 (BHL1+BHL2)/2	-211	-476%	4.688	6.684	11	0.270	11

Tick by tick effective data are used to generate the monthly effective spread.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 21: TAQ (Effective Spread) S&P 600 2014.01-2014.12

Effective Spread = 364 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking (RMSE)	Correlation	Ranking (Corr)
SHL1	-119	-253%	2.571	3.610	10	-0.026	11
SHL2∗ (negatives to be zero)	76	-39.87%	0.580	0.704	1	0.631	3
BHL1⋆ (mean spreads)	-135	-269%	2.613	3.750	12	-0.039	12
BHL2⋆ (mean parameters)	-117	-249%	2.557	3.569	9	-0.047	14
BHL3⋆ (negatives to be zero)	303	194%	1.830	2.665	7	**0.782**	1
CS1ˆ (mean spreads)	9	-138%	1.645	2.145	4	0.361	5
CS2ˆ (mean parameters)	-127	-270%	3.190	4.178	13	0.050	10
CS3‡ (negatives to be zero)	205	107%	1.286	1.675	2	0.606	4
ROLL	320	400%	4.091	5.718	14	0.316	6
AR‡ (negatives to be zero)	229	138%	1.751	2.231	5	**0.715**	2
Combination1 (BHL1+SHL2)/2	-30	-154%	1.410	2.089	3	0.160	8
Combination2 (SHL2+CS2)/2	-26	-155%	1.715	2.310	6	0.234	7
Combination3 (CS1+BHL1)/2	-63	-204%	2.089	2.921	8	0.097	9
Combination4 (BHL1+BHL2)/2	-126	-259%	2.557	3.639	11	-0.044	13

Tick by tick effective data are used to generate the monthly effective spread.

Definition of the abbreviations are given by Table 1

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Non-US equity markets applications

Figures 5 to 7 show the monthly average estimated spreads for equities listed on the London, Hong Kong and Thai stock exchanges respectively. Data was obtained from Bloomberg.

We observe several increases in bid-ask spreads estimated by SHL2, i.e. transaction costs, around notable market events. For example, the average spread jumped to over 0.8% when the sterling crisis occurred in September 1992 (Figure 5). When the Asian financial crisis began in July 1997, transaction costs rose significantly in both the Thai and Hong Kong equity markets (Figures 6 and 7). The collapse of Lehman Brothers in September 2008 and the financial crisis which heralded drove a jump in spreads in almost all equity markets used in our samples. After the results of the Brexit referendum became clear in June 2016, transaction costs in the UK equity market also appeared to rise sharply.

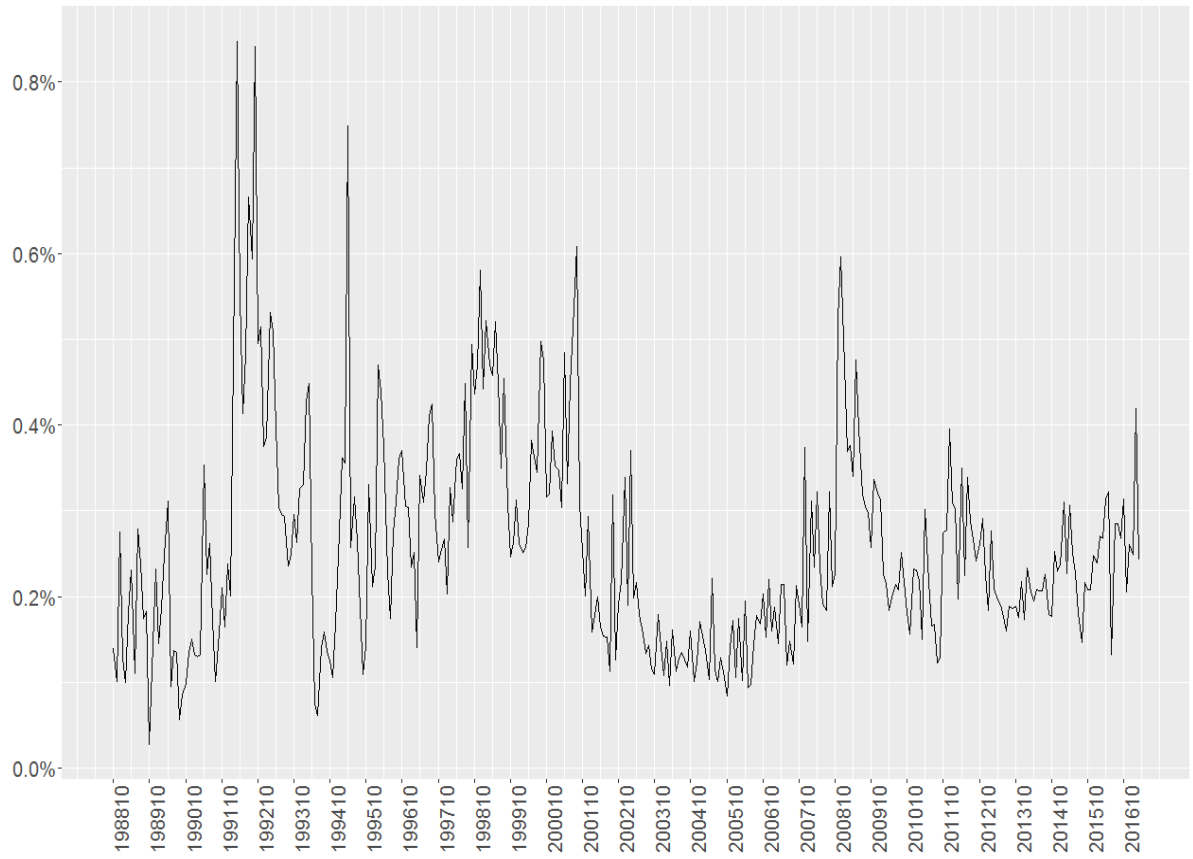

Figure 5: Monthly average estimated spread by SHL2 from 1990 to 2017, UK

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the London Stock Exchange on a monthly basis from October 1988 to March 2017. The figure plots the monthly equally weighted average spread of all stocks with at least 16 daily spread observations within the month. All data is taken from Bloomberg.

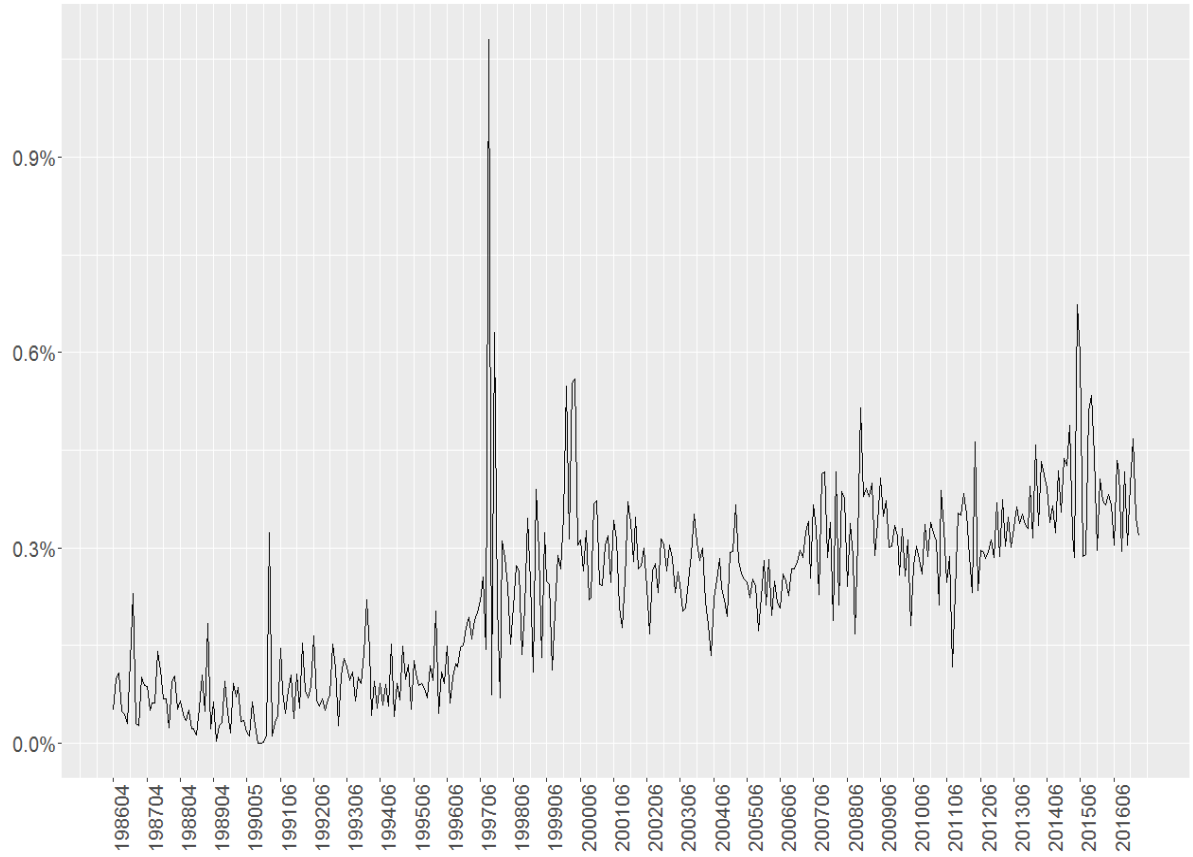

Figure 6: Monthly average estimated spread by SHL2 from 1986 to 2017, Hong Kong

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the Hong Kong Stock Exchange on a monthly basis from April 1986 to March 2017. The figure plots the monthly equally weighted average spread of all stocks with each recording at least 16 daily spread observations within the month. All data is taken from Bloomberg.

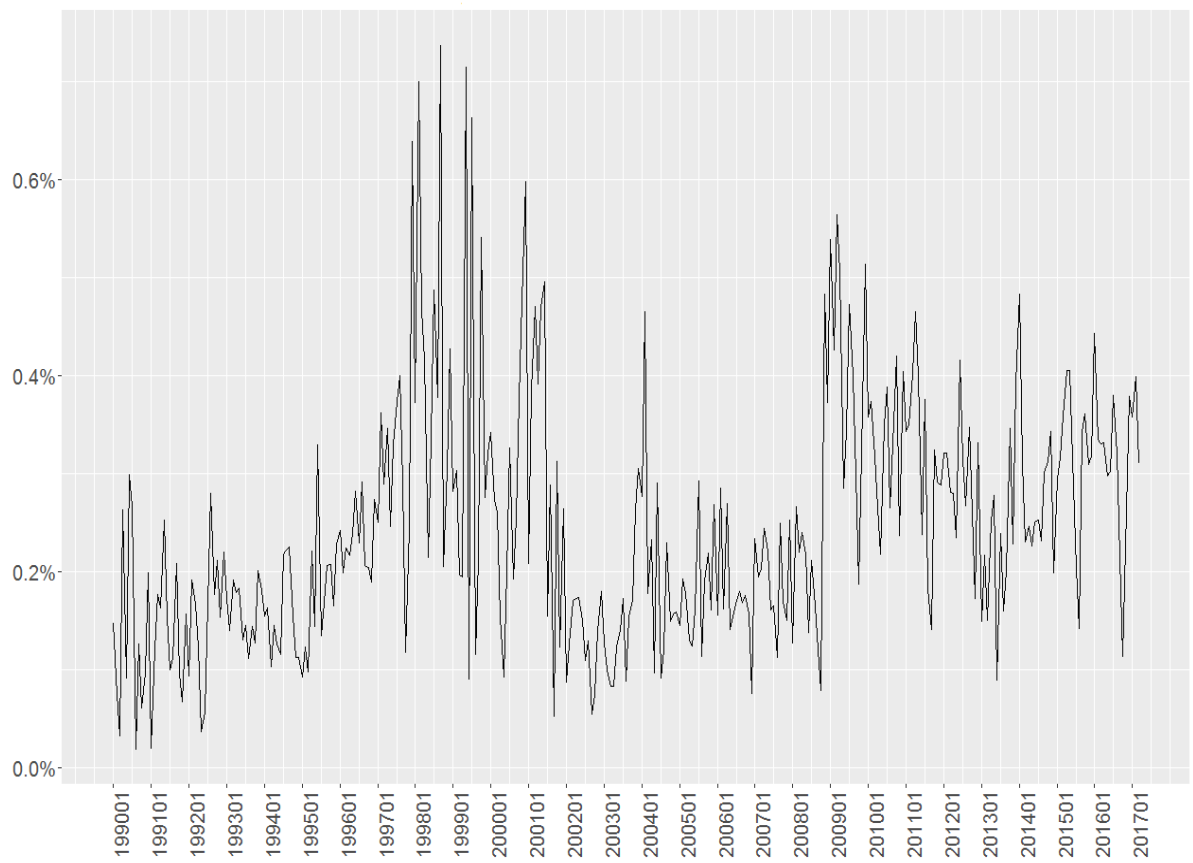

Figure 7: Monthly average estimated spread by SHL2 from 1990 to 2017, Thailand

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the Stock Exchange of Thailand on a monthly basis from January 1990 to March 2017. The figure plots the monthly equally weighted average spread of all stocks with each recording at least 16 daily spread observations within the month. All data is taken from Bloomberg.